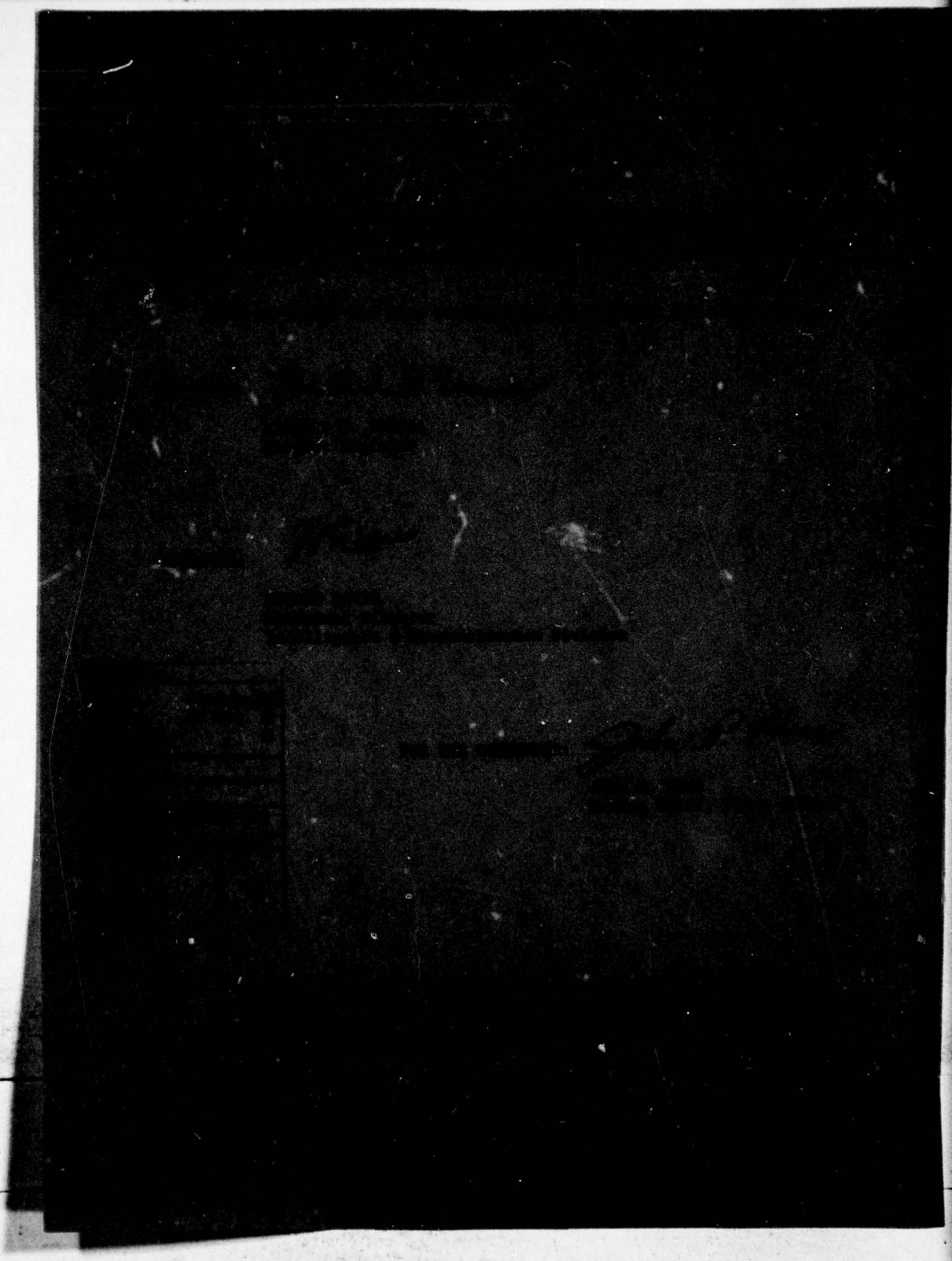
AD-A065 160 TEXAS INSTRUMENTS INC DALLAS F/G 17/2
TOTAL VOICE SPEAKER VERIFICATION.(U)
JAN 79 R L DAVIS, B M HYDRICK, G R DODDINGTON F30602-76-C-0329
UNCLASSIFIED RADC-TR-78-260 NL

1 OF 2
ADA
065160



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: RADC-TR-78-260	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): TOTAL VOICE SPEAKER VERIFICATION	5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report, July 1976 - May 1978
6. PERFORMING ORG. REPORT NUMBER: N/A	
7. AUTHOR(s): Robert L. Davis, Barbara M. Hydrick, George R. Doddington	8. CONTRACT OR GRANT NUMBER(s): F30602-76-C-0329
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Texas Instruments Incorporated, 13500 North Central Expressway, Dallas TX 75265	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 63714F, 681E0516, 17 05
11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRAA), Griffiss AFB NY 13441	12. REPORT DATE: January 1979
	13. NUMBER OF PAGES: 125 / 124 p.
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Same	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: Richard S. Vonusa (IRAA)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

speaker verification
voice authentication
digit recognition
word recognition
speech processing
pattern recognition
clustering

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The objective of this research has been to develop a robust speaker-independent, connected digit-sequence recognition capability as the front-end for a speaker verification (voice authentication) program and to install and demonstrate that capability on the Base and Installation Security System Advanced Development Model for speaker verification located at RADC. In such a system, the correct digit sequence recognition provides the user identification of the claimed identity. Verification is then performed on the same speech data.

(Cont'd)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Item 20 (Cont'd)

This total-voice system must recognize connected digits independent of speaker with high reliability. Two sequence constraints aid recognition: two parity checks must be satisfied, and "difficult" digit pairs are disallowed. A further sequence constraint added to aid verification was that all digits must be different. The selected constraints yield 320 possible sequences.

The speech processing strategy features highly reliable time registration and accommodates multiple concurrent hypotheses at various processing levels. Basic to robust speaker-independent recognition is the existence of a set of reference patterns capable of allowing for the speaker's sex and dialect. Rather than arbitrary segmentation of the design data to produce reference patterns, a hierarchical clustering algorithm was used, followed by an iterative optimization procedure. Good correspondence was found between the resulting clustered patterns and expected acoustic-phonetic distinctions such as sex, context, and dialect.

Tests were run on the speaker-independent connected digit sequence recognition system. One test used a data base of 1060 six-digit sequences from 106 speakers (64 males, 42 females). Samples of 100 of the 320 sequences appeared. No sequence was found (rejection) for 1.0 percent of the sequences; an incorrect sequence was found (substitution), for 0.5 percent of the sequences. A second test was run on all 320 sequences said by 10 subjects, plus 195 of the 320 sequences said by an eleventh subject, yielding a 6.0-percent rejection rate and a 3.4-percent substitution rate.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

EVALUATION

The objective of this effort was to develop a speaker independent digit sequence recognition front-end to a speaker verification system. The Total Voice concept eliminates the need for badge readers, keyboards, etc. for inputting the users' claimed identity. The system allows an individual to speak only his code number, such as his Social Security Number, work badge number, etc. The technique automatically recognizes the digit code independent of speaker, and then uses the same acoustic data to verify the individual.

This effort demonstrated that the Total Voice concept is a successful means of increasing user throughput into a secure area.

The concept was implemented on the ESD/RADC Advanced Development Model Speaker Verification System and is demonstratable at RADC's Entry Control Laboratory.

More work is needed in the area of improving the speaker-independent connected digit recognition algorithms and integrating them into an operational Total Voice System.

As a result of this successful implementation of the Total Voice concept, future plans call for continuing technology development for an operational environment.

Richard S. Vonusa

RICHARD S. VONUSA
Project Engineer

SECTION I
INTRODUCTION

A. BACKGROUND

This final report covers the fifth in a series of programs undertaken by Texas Instruments to further develop speaker verification (voice authentication[1,2]) technology. In the first program,[3] a promising high-performance speaker verification technology was developed and comprehensively tested in a laboratory environment, with accurate and reliable methods of time registration providing a major performance impact.

In the second program,[4] operationally important problems were solved to provide an operational capability for applications such as automatic entry control. Concurrent with this second program were:

The development of an Advanced Development Model voice verification system for the Base and Installation Security Systems (BISS) program under Electronic Systems Division sponsorship[5]

The installation of an operational, fully automated entry control system, internally funded, to provide entry control to the Texas Instruments Corporate Information Center.[6]

In the third program,[7] advanced speech processing capabilities were developed to enhance speaker verification effectiveness and extension of speaker verification technology was made to other applications. Effort was focused on two specific applications: speaker verification using passwords embedded in free text and speaker identification (and subsequent verification) using spoken identification codes (called "Total Voice" verification). Both of these required major emphasis on the development of word recognition technology and the integration of recognition and verification techniques.

The fourth program[8] was a study conducted to develop speaker verification techniques for use over degraded communication channels–specifically telephone lines. A test of BISS type speaker verification technology was performed on a degraded channel and compensation techniques were then developed.

This fifth program was the coalescence of the Total Voice verification technology and the hardware of the Advanced Development Model BISS speaker verification system, shown in Figure 1, culminating in the installation on the BISS-SV system, of the Total Voice computer program, called TVBISS.

The remainder of this section discusses the concept of Total Voice in more detail. Section II reviews the speech processing used at Texas Instruments, and specifically on this program. Section III briefly describes the data sets collected, and Section IV discusses the clustering methods developed under this contract for use in creating a set of speaker-independent reference patterns used in digit sequence recognition. Section IV also includes a very detailed discussion of the resulting reference patterns.

Figure 1. BISS System

A wide flexibility exists in the selection of various parameters and thresholds in the Total Voice computer program. The tradeoff testing done during these selections as well as the final results of digit sequence recognition tests are described in Section V. Section VI includes both a description of the speaker verification part of the program and the results of a limited speaker verification test run using the final program. Conclusions and recommendations are given in Section VII. The Appendix of this report is a paper presented at the *Fourth International Joint Conference on Pattern Recognition* covering the work on this contract. It is suggested that the reader desiring only a summary discussion of the contents of this report would be advised to read this Appendix instead of the full report.

B. TOTAL VOICE

Two key functions are provided by the user in an entry control system. These functions are: user identification and user verification. The verification function is required to be performed on a personal attribute; in this case, on the user's voice characteristics. Existing techniques for user identification are manual and are based upon badge or keyboard identification entry. It is desired to eliminate the manual identification mode and consolidate identification and verification. This is done by using a *spoken* identification code. Two benefits accrue from a Total Voice speaker verification capability: first, verification time is reduced considerably. This is possible because the input speech data used for identification may also be used for verification, thus completely eliminating the speech input time required for verification. Second, eliminating all but speech input provides operational advantages. Hands need not be freed to operate manual identification devices, and the verification terminal becomes less expensive and more mobile.

It is important to note that the consolidation of user identification and voice verification is intended to be distinct from the problem of speaker recognition; i.e., the identification of the user is based upon a unique identification code assigned to that user and not upon the unique properties of his voice. If identification were based upon personal voice characteristics, then identification performance would deteriorate rapidly with increasing population size. On the contrary, with identification based on a unique identification code assigned to each user, the identification performance does not deteriorate rapidly with increasing population size. Identification performance in this case is determined by identification code uniqueness, which may be increased rather arbitrarily through various methods of adding redundancy.

The next question concerns how to recognize the spoken identification code. Perhaps the most straightforward method is to recognize that code by comparison with user-specific speech reference data. This approach avoids the problem of recognizing speech independent of speaker, but is not viable because required processing is directly proportional to the number of users and, therefore, identification processing becomes prohibitively expensive with even modest population sizes. The approach used here for recognizing the identification code is to represent the code as a sequence of words selected from a small set of words (the 10 digits) and to recognize the words composing the identification code by speaker-independent word recognition. Of course, there are problems in performing speaker-independent word recognition that are exacerbated by the application. In this case, there exists the requirement for minimal user training. Minimal user training implies the use of normal continuous speech input. A discrete input code identification scheme would suffer seriously from user violation of the discrete speech requirements. The ability to handle connected speech is, therefore, a must. At the same time, however, reliable

speaker-independent recognition of these digit sequences said in continuous speech is necessary. This recognition performance is aided by incorporating three constraints into the sequence recognition:

Two parity checks must be satisfied

"Difficult" digit pairs are disallowed

All digits must be different.

The minimum desired number of identification codes for this contract was 300. Applying the above constraints, 320 codes are obtained using a code length of six digits.

After a six-digit sequence is recognized, user verification is done on the same input data used for the sequence recognition. The recognized six-digit sequence is used to locate the reference file for that claimed identity. (The verification presumes the prior enrollment of each speaker on the system.) The reference file is then used to make the verification decision.

An overall flowchart for the digit recognition and verification process is shown in Figure 2. More details concerning the sequence recognition and the user verification are presented in this report.

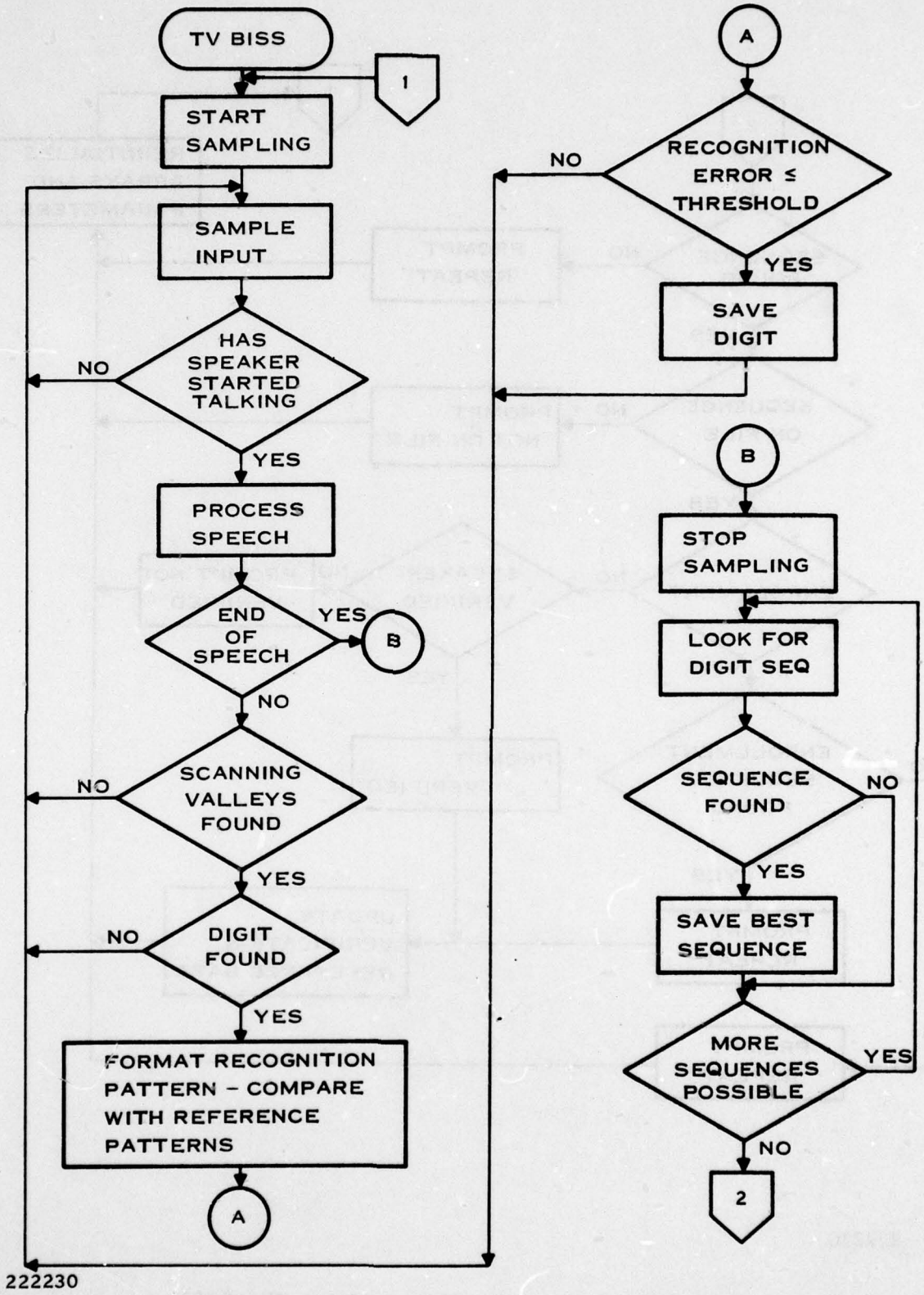
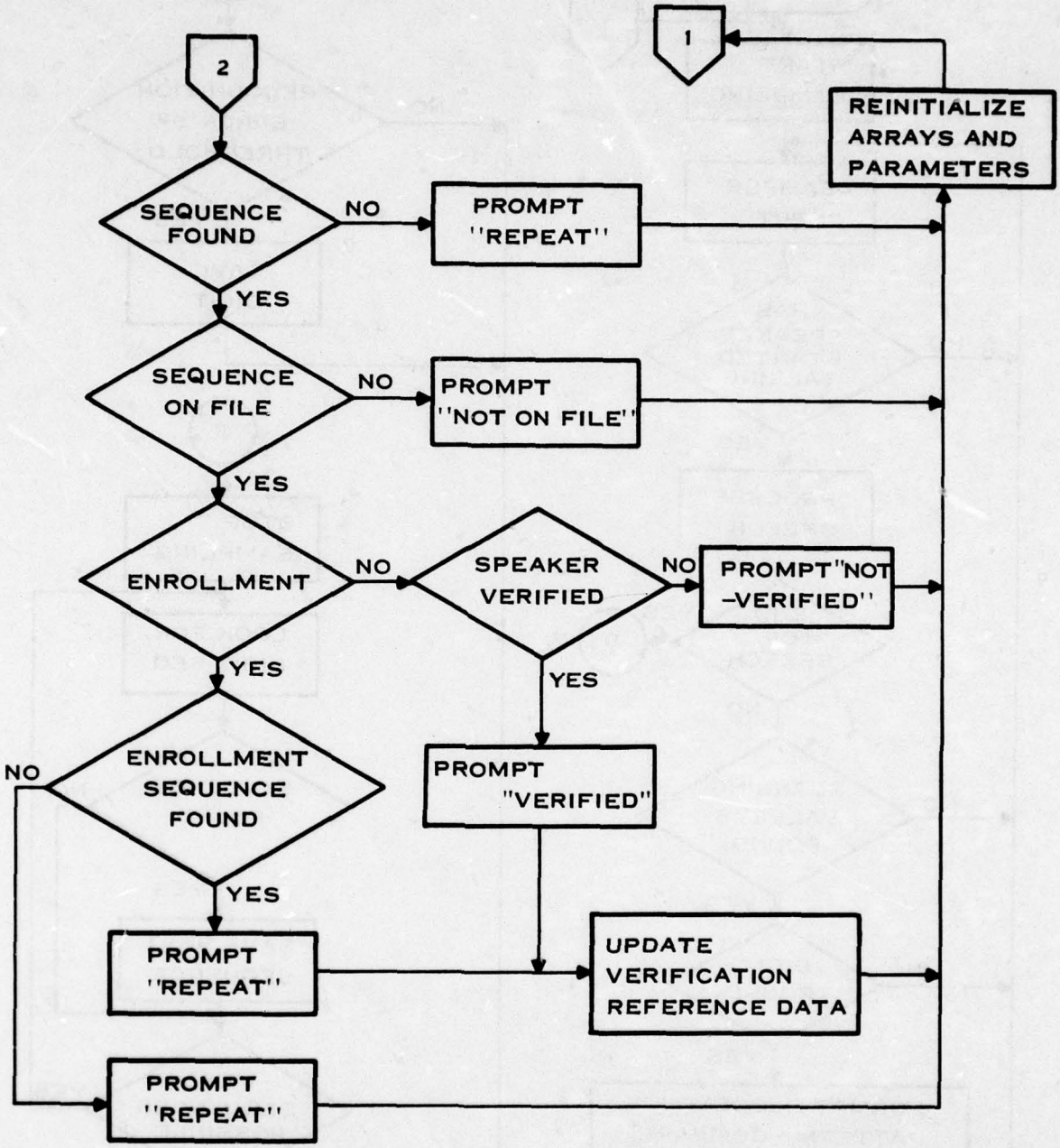


Figure 2. Digit Recognition and Verification Process (Sheet 1 of 2)



222230

Figure 2. Digit Recognition and Verification Process (Sheet 2 of 2)

SECTION II
SPEECH PROCESSING

A. SPECTRAL PROCESSING

The speech processing strategy used in this program is based upon the relative spectrum of speech as a function of time, which is the output of a 16-channel digital filter bank that has been preprocessed as described in this section.

1. Filter Bank Definition

The spectrum is obtained by processing the speech signal through a digital filter bank preceded by a first order differencing network (for preemphasis). The filter bank consists of 16 bandpass filters, each followed by a fullwave rectifier and a four-pole lowpass Bessel filter with a 3-dB cutoff at 30 Hz. Each of the 16 filters is sampled 100 times per second. The digital filter characteristics are given in Table 1 and a block diagram of the spectral analysis hardware is shown in Figure 3. Actual filter responses appear in Figure 4 for the bandpass filters alone and in Figure 5 for the bandpass filters with preemphasis.

TABLE 1. CHARACTERISTICS OF 16-CHANNEL FILTER BANK

Filter	Center Frequency (Hz)	Bandwidth (Hz, at −6 dB)
1	280	250
2	395	280
3	525	310
4	630	340
5	750	360
6	900	360
7	1080	360
8	1265	365
9	1480	365
10	1725	365
11	1985	365
12	2285	360
13	2640	365
14	3150	625
15	3720	635
16	4235	615

For processing, the top three filters are summed and filter 14 is replaced by this sum. Filters 15 and 16 are set to zero. The resulting 14 filter outputs at each time sample are represented by the spectrum amplitude vector:

$$A_j = \begin{Bmatrix} a_{1j} \\ a_{2j} \\ . \\ . \\ . \\ a_{14,j} \end{Bmatrix} = \begin{Bmatrix} a_1(t_j) \\ a_2(t_j) \\ . \\ . \\ . \\ a_{14}(t_j) \end{Bmatrix}$$

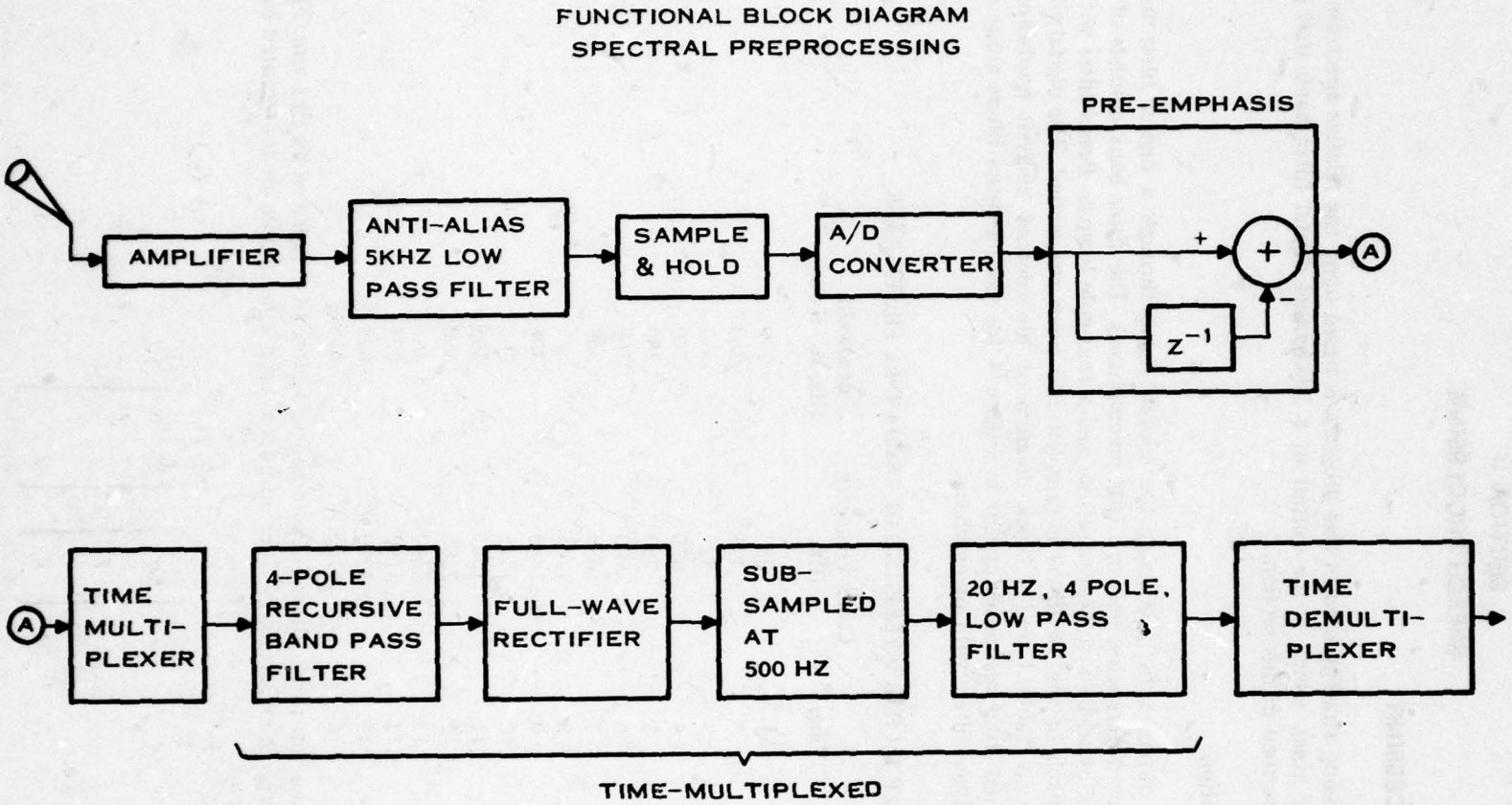


Figure 3. Functional Block Diagram of Spectral Preprocessing

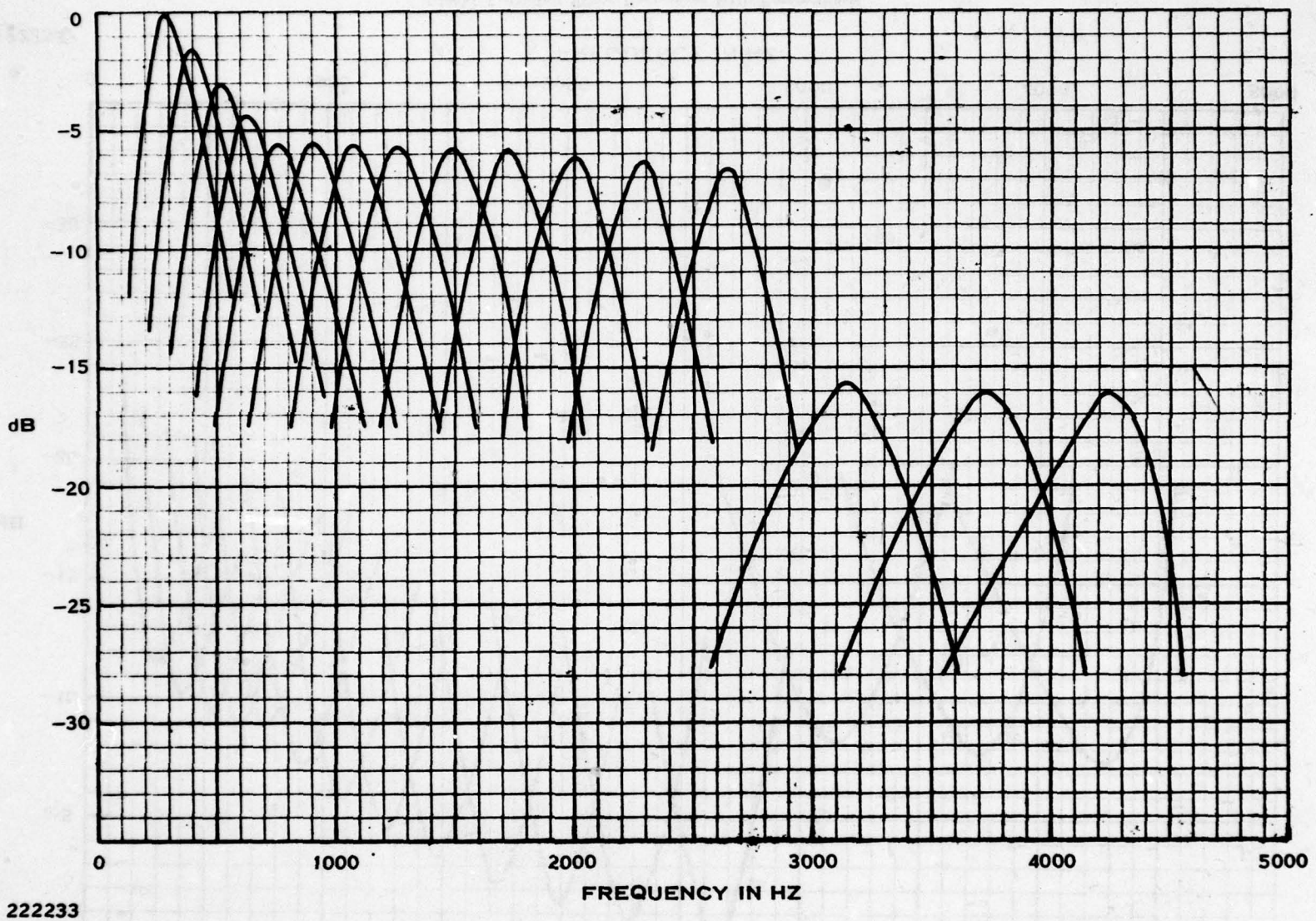


Figure 4. Digital Filter Responses

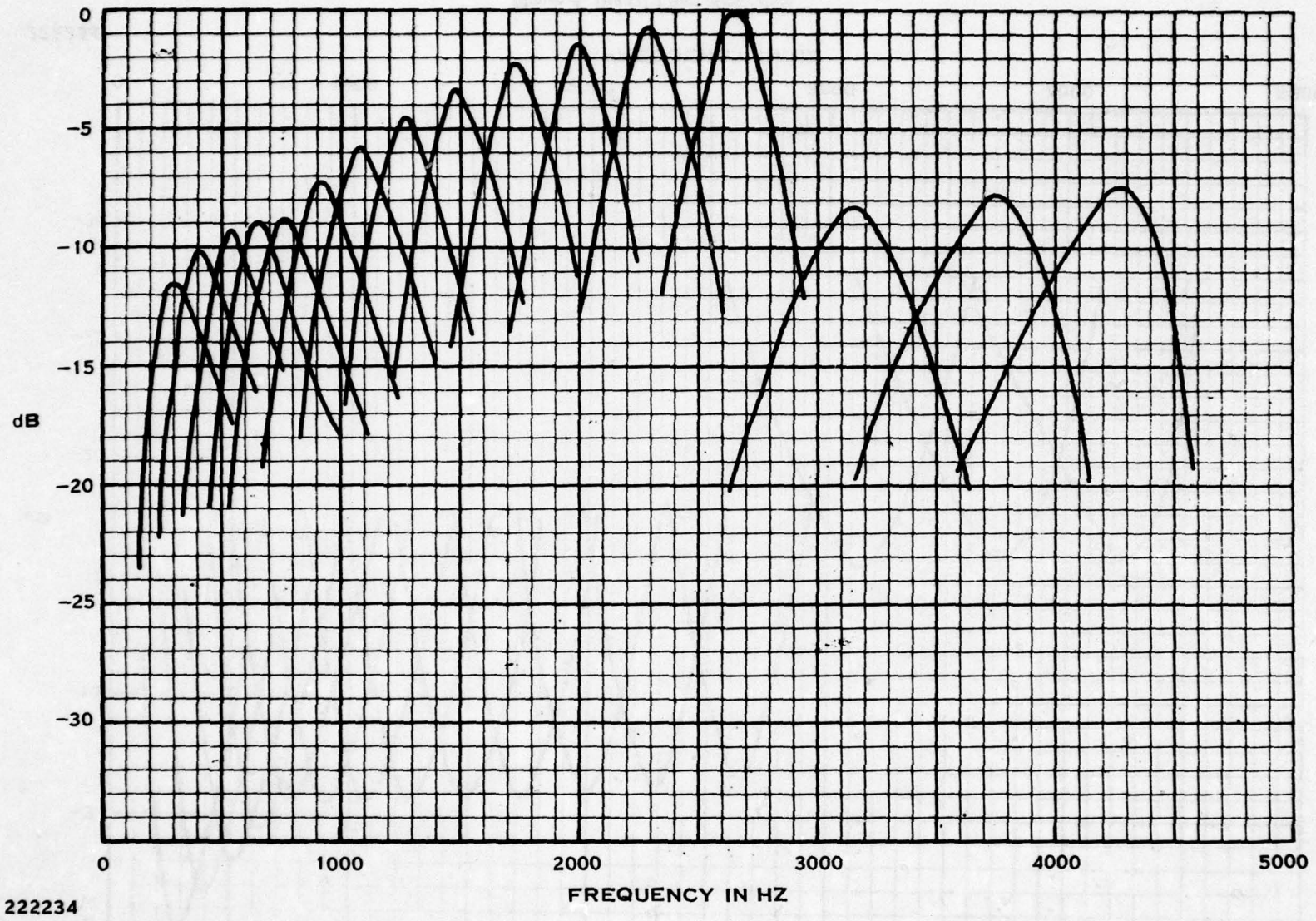


Figure 5. Digital Filter Responses With Preemphasis

2. Regression

It has been found that by eliminating the gross aspects of the spectrum, such as the slope and curvature, more clearly defined formant frequencies are obtained.[7] Therefore, the spectrum amplitude vector is regressed by the first three elements of an orthonormal basis set:

$$(A_j)_R = A_j - \sum_{k=0}^{2} c_{jk} F_k$$

where

$$F_k = \begin{Bmatrix} f_{1k} \\ . \\ . \\ . \\ f_{14,k} \end{Bmatrix} \quad k = \{0, 1, 2\}$$

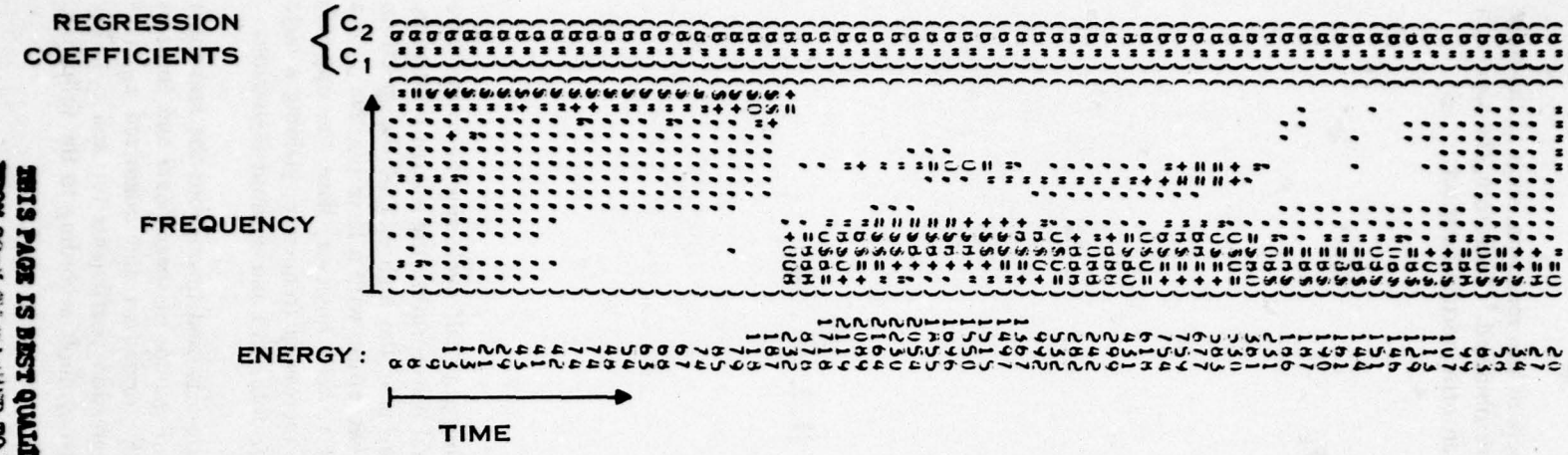
$$\left.\begin{aligned} f_{i0} &= \frac{1}{\sqrt{14}} \\ f_{i1} &= -\frac{1}{\sqrt{7}} \sin\left[\frac{(i - 1/2)}{14}\ \pi\right] \\ f_{i2} &= -\frac{1}{\sqrt{7}} \cos\left[\frac{(i - 1/2)}{14}\ \pi\right] \end{aligned}\right\} \quad i = \{1, 2, \ldots, 14\}$$

and

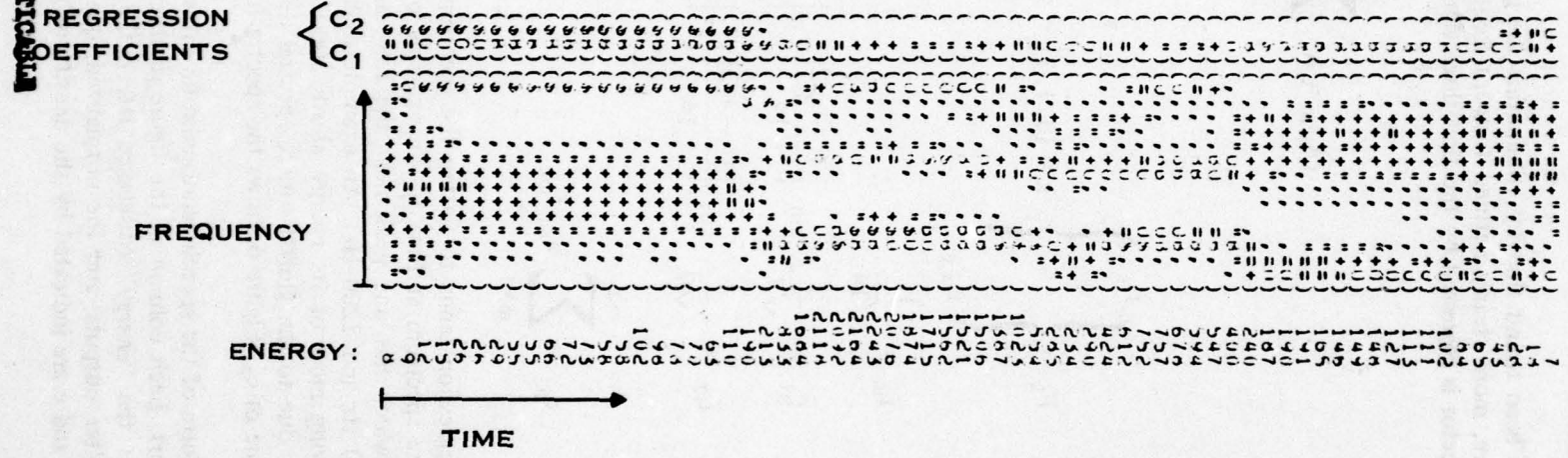
$$c_{jk} = \sum_{m=1}^{14} a_{mj} f_{mk}$$

Thus, the regression tends to flatten the spectrum, removing any half cycle sine or cosine wave trends of the spectrum at time t_j. An example of a spectral waveform having a large positive c_1 is a nasal, which has one peak near the low end and one near the high end of the spectrum (around 250 Hz and 2200 Hz). An example of a spectral waveform with a large positive c_2 is a sibilant, having most of its energy above 3000 Hz. Most vowels, however, have the opposite spectral tilt due to the glottal source spectral decay with increasing frequency, yielding a large negative value of c_2. Figure 6 shows the spectra for /sɛvən/, both with and without regression.

The form of the speech representation shown in Figure 5 is used throughout the remainder of this report. Each column in the figure represents 10 milliseconds of speech data and contains 17 features: the "energy" (discussed later); the regressed, normalized and quantized a_{ij}, (i = 1,. . .,14) filter outputs; and the normalized, quantized regression coefficients (c_1 and c_2). The value of a_{ij} and c are indicated by the density of the printed symbols according to the following:



(A) SPECTROGRAM OF PREPROCESSED SPEECH WITHOUT REGRESSION FOR "SEVEN"



(B) SPECTROGRAM OF PREPROCESSED SPEECH WITH REGRESSION FOR "SEVEN"

222235

Figure 6. Comparison of Pre/Postregression Spectra for "Seven"

Value	0	1	2	3	4	5	6	7
Symbol	blank	,	"	+	=	0	B	$

At this point, the energy is not quantized; however, it is always used relative to other energies and the relative value is then quantized.

Note from Figure 6 that the normalized, quantized values of c_1 and c_2 for no regression are 2(") and 6(B), respectively. Values above these represent positive c and below, represent negative c. For example, c_2 during the vowel in the regressed spectra in Figure 6, is zero (blank) indicating the removal of much of the downward spectral tilt with increasing frequency.

Also, note that the nonregressed spectra in Figure 6 appears worse than it should since the quantization (discussed later in this section) thresholds were chosen using a regressed data set. A truer representation would require recomputation of the quantization thresholds using non-regressed data.

3. Normalization

The regressed amplitude vector is next normalized by a modified postregression standard deviation, σ_j^* for time t_j:

$$\sigma_j^* = \sigma_{post_j} + \sigma_{min}$$

where

$$\sigma^2_{post_j} = \frac{1}{11}\left(\sum_{m=1}^{14} a_{mj}{}^2 - \sum_{k=0}^{2} c_{jk}^2\right)$$

$$\sigma_{min} = 62 \text{ for Total Voice}$$

However, it has been noticed that regression sometimes eliminates too much of the variance of the filter output vector $\bar{A}_j$. To limit the regression, a limit is placed on σ_{post}, as follows

$$\sigma^2_{post_j} = \max\left(\sigma^2_{post_j}, R^2_{min}\,\sigma^2_{pre_j}\right)$$

where

$$\sigma^2_{pre_j} = \frac{1}{13}\left(\sum_{m=1}^{14} a^2_{mj} - c^2_{jo}\right)$$

$$R_{min} = 0.6$$

Note, that when $\sigma_{post_j} = R_{min}\,\sigma_{pre_j}$, the regression coefficients c_1 and c_2 are reduced in order to decrease the amount of regression.

The resulting normalized amplitude vector is then:

$$(A_j) = \frac{1}{\sigma_j^*} (A_j)_R$$

The regression coefficients c_{j1} and c_{j2} are also normalized by σ_j^*.

4. Quantization

The regressed and normalized amplitude vector is then quantized to one of eight levels according to a set of quantization thresholds ϕ_{iq} :

$$(a_{ij})_Q = q \text{ IFF } \begin{cases} (a_{ij})_N \geq \phi_{iq} \\ (a_{ij})_N < \phi_{i,q+1} \text{ for } q = 0, 1, \ldots, 7 \end{cases}$$

where $\phi_{iq} < \phi_{i,q+1}$; $\phi_{i0} = -\infty$; and $\phi_{i8} = \infty$.

Rather than have these quantization levels (ϕ_{iq}) being chosen to yield a uniform probability, however, it was felt to be more desirable to have the quantization thresholds cluster at higher energy levels (p. 16 of Reference 4). In this way the sensitivity to noise can be reduced and quantization resolution is increased in the region of interest (which is the spectrum amplitude at the formant frequencies).

The quantization thresholds were chosen by plotting histograms for each of the regressed, normalized filter outputs [$(a_{ij})_N$'s] during regions of formant locations for each of the vowels in a specially collected design data set. This data set was one repetition by each of 10 males and 10 females of the set of words given in Table 2.

TABLE 2. WORDS USED IN DETERMINING QUANTIZATION THRESHOLDS

*Pot	*Bert	*Bet	*Bought
*Put	Bout	Bait	*Beet
Boyd	*Bat	Boat	*But
Butte	Bite	*Bit	*Boot

*Pure vowels

In actually computing quantization thresholds, only the pure vowels (*in Table 2) were used. The filters for each of the vowels used is given in Table 3, along with the expected formant locations for each vowel, as given in Peterson and Barney.[9]

A plot of the octiles for each of the 14 filters is given in Figure 7. The top and bottom curves were each replaced by the constant + sine + cosine best fit to smooth the curves. The

TABLE 3. FILTER NUMBER OUTPUTS USED IN DETERMINING QUANTIZATION THRESHOLDS

	Formant Locations–Males				Formant Locations–Females			
	Design Data Energy Peaks	Peterson and Barney F_1	F_2	F_3	Design Data Energy Peaks	Peterson and Barney F_1	F_2	F_3
/ɑ/	5–7, 13–14	5	7	12–13	6–7, 13–14	5–6	8	13
/U/	2–3, 8–9, 12	2	6–7	12	3, 10, 13	2–3	7–8	13
/b/	2–3, 8–10	2–3	8–9	10	3–4, 9–11	2–3	9–10	11
/æ/	3–5, 9–10, 13	4	10	12–13	5, 11–12	5–6	11	13
/ɛ/	3–4, 10–11, 13	3	10–11	12–13	3–5, 11–12	3–4	12	13–14
/I/	2–3, 10–11, 13	2	11	13	2–3, 11–13	2	12–13	14
/ɔ/	3–5, 12–14	3	5–6	12–13	4–6, 13–14	3–4	6	13
/i/	1–2, 11–13	1	12	14	1–3, 13	1	13	14
/Λ/	3–4, 8–9, 12–13	4	7–8	12	4–5, 10–11, 13	5	9	13
/u/	1–2, 9, 12	1	6	12	2–3, 10–11, 13	1–2	6	13

middle five curves were then replaced by linear interpolations between the top and bottom one to yield Figure 8, which gives the quantization coefficients for the 14 filters.

The octile ranges for c_1 and c_2 were also determined from histograms of the normalized c_1 and c_2 values except the limiting of the $\sigma_{post_j} \geq R_{min}\ \sigma_{pre_j}$ as described in the previous section was not done. This yielded the following values of ϕ_q for c_1 and c_2.

	ϕ_0	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	ϕ_8
C_1	$-\infty$	−3.0	−1.5	0	1.5	3.0	4.5	6.0	∞
C_2	$-\infty$	−7.0	−5.67	−4.33	−3	−1.67	−.33	1.0	∞

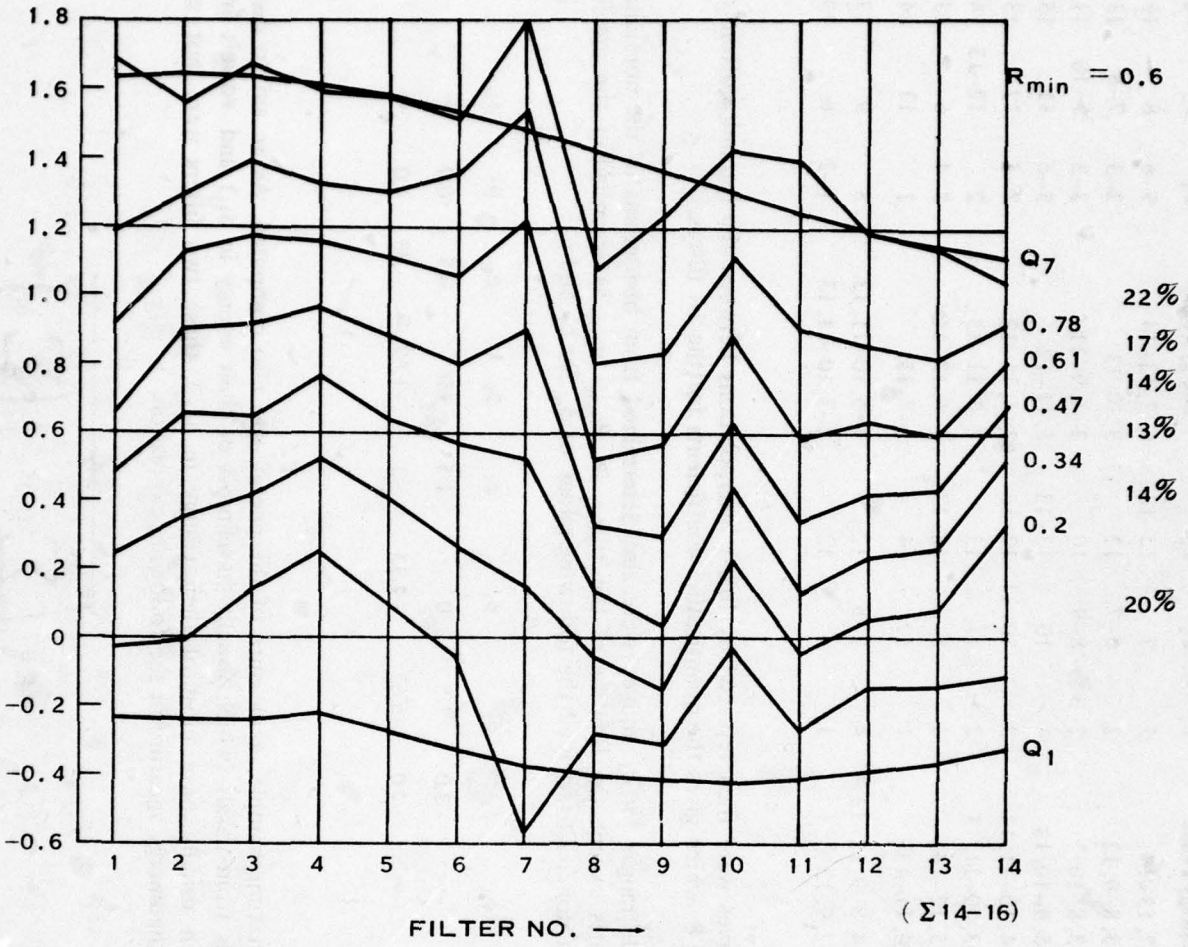
5. Energy

For each time sample, a measure of the energy was also computed. As an aid to distinguishing vowels from nasals (which usually have most of their energy in a_{1j}) and vowels from sibilants (which usually have most of their energy in a_{14j}), these two filters were not used in computing the energy measure in the following expression.

$$E = \sqrt{\sum_{i=2}^{13} (a_i)^2 - \frac{1}{11}\left(\sum_{i=2}^{13} a_i\right)^2}$$

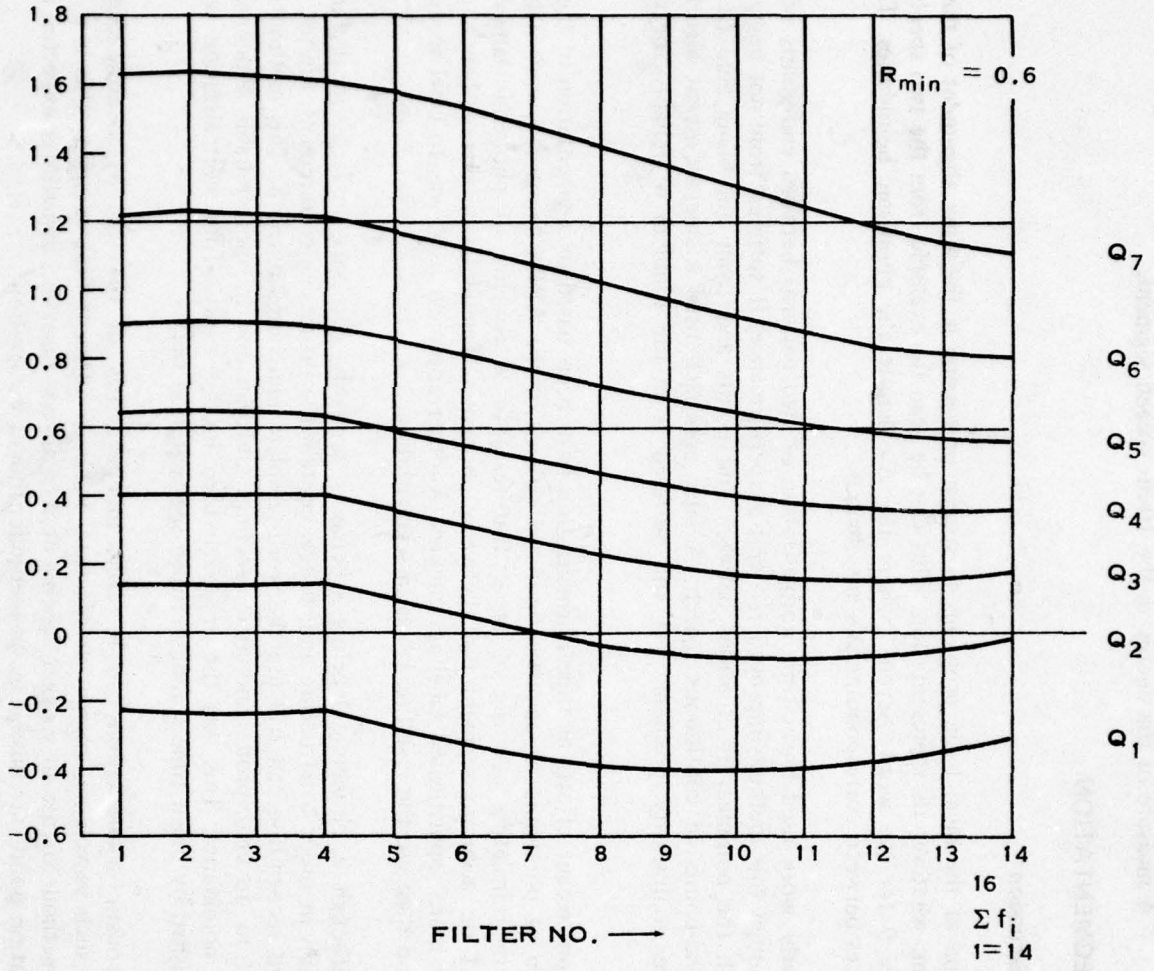
6. The Preprocessed Speech Representation

In summary, the input speech representation used is a 17-element vector representing a 10-ms segment of speech. The vector comprises:



222236

Figure 7. Octiles for Filter Outputs for Determining Quantization Thresholds



222237

Figure 8. Quantization Thresholds for Filters

Fourteen outputs (a_i) of a digital filter bank that have been regressed, normalized, and quantized to 3 bits each

Two regression coefficients (c_1 and c_2) that have been normalized and quantized to 3 bits each

A measure of the energy in the 10-ms speech segment.

B. SEGMENTATION

1. Discussion

One of the most basic problems in speech processing is the time alignment of the speech waveform, whatever its representation. This can be seen for example from the two spectrograms in Figure 9 for the word "seven," where the Δ's denote the phonemic boundaries. The time differences between corresponding Δ's are obvious.

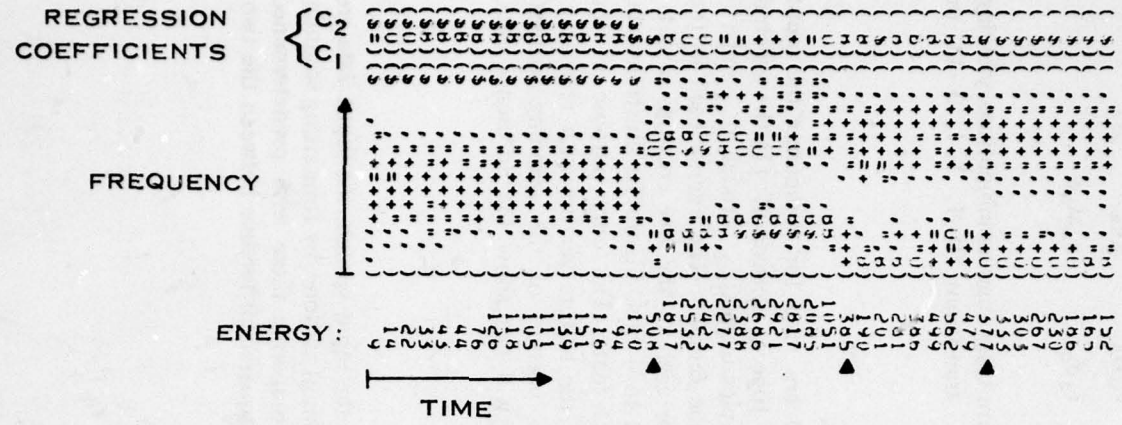
Early work used linear time normalization of two patterns between end-points of words, and although this method improved recognition performance, it suffered from not being able to deal with the nonlinear fluctuations between end points and from not being able to reliably locate end-points in continuous speech. A later approach used a steepest-ascent algorithm to solve the nonlinearity problem, but still suffered from the problem of reliably locating end points.

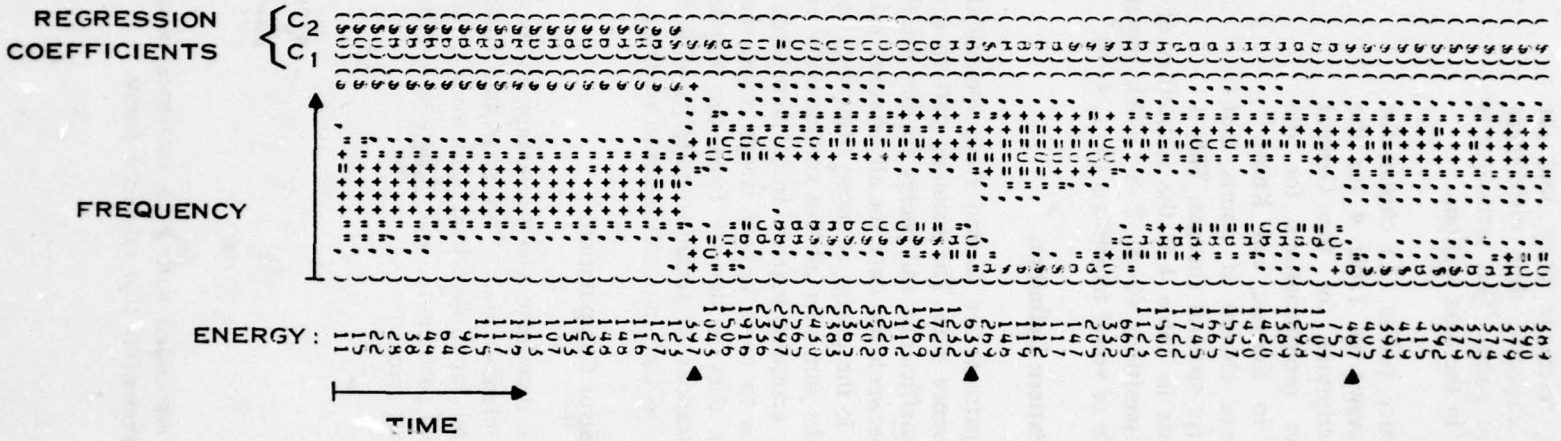
More recent efforts at time normalization have been based on segmentation of the speech waveform at phonemic or acoustic boundaries. Most of the ARPA sponsored work (Reddy[10]) dealt with translating a string of input features into a sequence of phonemic labels, which depended on accurate segmentation between phonemes. Segmentation errors then had to be fixed by more sophisticated labeling schemes. Another approach used was to label speech every 10 ms and then use the resulting labels in segmentation.

Although a phonemic labeling scheme is probably necessary in large vocabulary word recognition in order to structure and, hence, contain the processing requirement, limited vocabulary word recognition can still use the word-template matching approach. This method obviates the need to do error-prone phonemic labelings and can use the more reliable segmentation at acoustic boundaries. This was the approach used in this study. Time normalization was then accomplished by linear time scaling between segmentation points.

Acoustic segmentation can be done independently of the text by locating changes in features such as voicing, energy, or spectrum between adjacent time samples. This is a reliable, precise method for use in speaker-dependent recognizers; however, sometimes expected acoustic segmentation points are missed in speaker-independent recognition.

The more robust approach used in this study is to use a text-dependent approach, matching a feature vector (called a scanning pattern) extracted from the input speech waveform to reference scanning patterns, or templates, computing a distance between the input and all reference patterns. Minima in this distance function then are locations of potential acoustic segmentation points (called reference points). Reference points are then combined into optimal sequences using a dynamic programming routine that accounts for the value of the distance





222238

Figure 9. Two Examples of Segmentation of "Seven"

function at the reference point and for deviations from expected time differences between reference points. This technique is described further in the next sections.

TABLE 4. REFERENCE POINT DEFINITION FOR THE DIGITS

$z_{\Delta}Ir_{\Delta}o_{\Delta}$	$f_{\Delta}aI_{\Delta}v$
$w_{\Delta}\Lambda_{\Delta}n$	$s_{\Delta}I_{\Delta}k_{\Delta}s$
$_{\Delta}t_{\Delta}u_{\Delta}$	$s_{\Delta}\varepsilon_{\Delta}v\partial_{\Delta}n$
$\theta_{\Delta}ri_{\Delta}$	$_{\Delta}e_{\Delta}t$
$f_{\Delta}o_{\Delta}r$	$n_{\Delta}aI_{\Delta}n$

The reference points (Δ) chosen for the digits are shown in Table 4 for the phonetic transcriptions for the General American dialect pronunciations for the digits as found in Kenyon and Knott.[11] These points were chosen at points that would exhibit large spectral changes. This is true of the points in Table 4 if the restriction of not permitting digit combinations yielding vowel-to-vowel transitions (0–8, 2–8, 3–8), vowel-to-glide transitions (0–1, 2–1, 3–1), or semivowel-to-glide or vowel transitions (4–1, 4–8).

2. Scanning Pattern Definition

Scanning patterns are formed from spectral data and are used for comparing the input speech with reference data. The scanning pattern formed at time t_j consists of: (1) the spectral data, regression coefficients, and energy for the five time samples from t_{j-2} through t_{j+2} and (2) the difference between the data for all adjacent pairs of time samples. The energy used in the scanning pattern is the energy measure (described earlier) for each of the five columns of data, normalized by the sum over all five columns and quantized to 4 bits. Figure 10 illustrates the formation of a scanning pattern from preprocessed speech data. The only purpose of the difference data is to more heavily weight rapid changes of the feature vectors with respect to time. Since this data is derived from the standard data portion of the scanning pattern, subsequent illustrations of scanning patterns in this report will not show the difference data, even though it is, in fact, part of the actual pattern.

3. Scanning Error Computation

In order to determine where reference points occur in the input speech, the input data are compared with reference data. This procedure (called scanning) is done by formatting scanning patterns from the input speech at each time sample t_j, comparing these with predetermined reference scanning patterns $\vec{r}_k$, and obtaining a measure of squared difference between the two, called the scanning error:

$$e_{kj} = ||\vec{x}_j - \vec{r}_k||^2 = \sum_{i=1}^{164} (x_{ij} - r_{ik})^2$$

The final error associated with each reference point is the minimum error of all comparisons with patterns representing that reference point.

4. Valley Finding

Using the scanning errors as a function of time, an error function is thus generated for each type of reference scanning pattern using the minimum scanning error for each pattern type for each time sample. (Multiple reference scanning patterns are allowed for each reference point of each digit.) Each function is monitored for dips of sufficient magnitude to be considered as potential locations of the corresponding reference points in the input data. These dips are called valley points when the ratio of the scanning error following the dip to the scanning error at the dip itself is greater than or equal to a specified peak-to-valley ratio (PVR) (typically 1.1 to 1.3), and the magnitude of the scanning error for the valley point is less than or equal to a threshold (typically 600 to 1,200). The occurrence of a peak (verified when the ratio of the scanning error following the peak to the scanning error at the peak is less than the reciprocal of the PVR) is required before another valley point can be found. The valley finding procedure is shown in Figure 11.

C. GENERATING WORD HYPOTHESES

The next task is to fit the hypothesized reference points together to form word hypotheses. A sequence of time-ordered reference-point hypotheses for a word must exist. The time distance between each pair of reference points must satisfy the word-specific minimum/maximum restrictions given in a later section. An error is determined for each reference point pair which is weighted by deviations from the expected distance between the two points and the scanning error at each hypothesized reference point. The weighted error for reference points i and i + 1 is:

$$E_{wi} = \frac{(e_i + \text{offset})\,(e_{i+1} + \text{offset})}{1024}\left[1 + \beta\left(\frac{\hat{dt}_i - dt_i}{\hat{dt}_i^*}\right)^2\right]$$

where

$$dt_i = t_{i+1} - t_i$$

$$\hat{dt}_i = \text{expected } dt_i$$

$$\hat{dt}_i^* = \max\,(\hat{dt}_i, \hat{dt}_{min})$$

$$\beta = 2$$

$$\hat{dt}_{min} = 4$$

$$\text{offset} = 100$$

$$e_i = \text{valley point error for reference point i}$$

The hypothesized word sequence error (SQ) is the sum of the E_w for all reference point pairs in the word and is limited by the word specific thresholds in the following list:

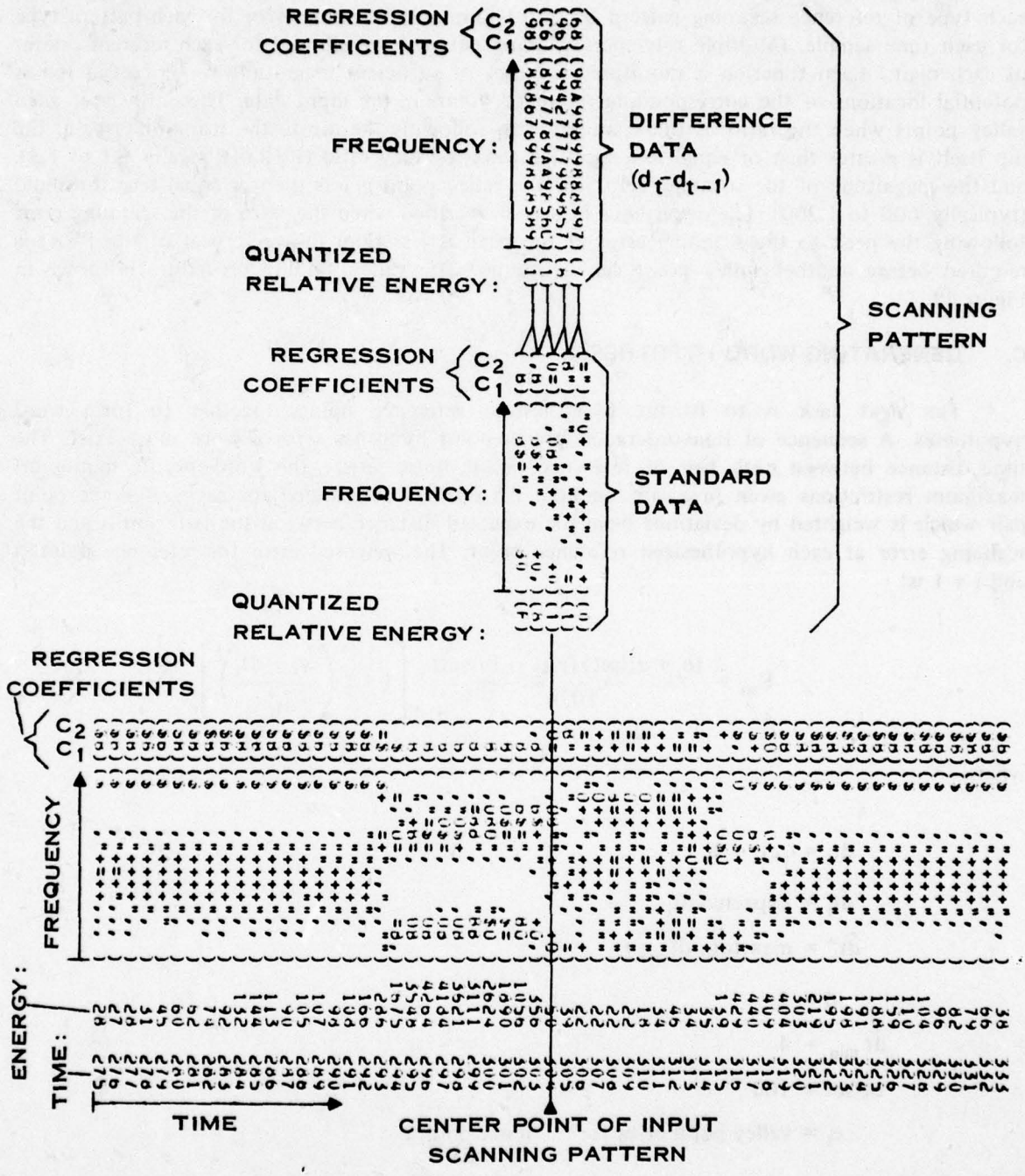


Figure 10. Example of Scanning Pattern Formation

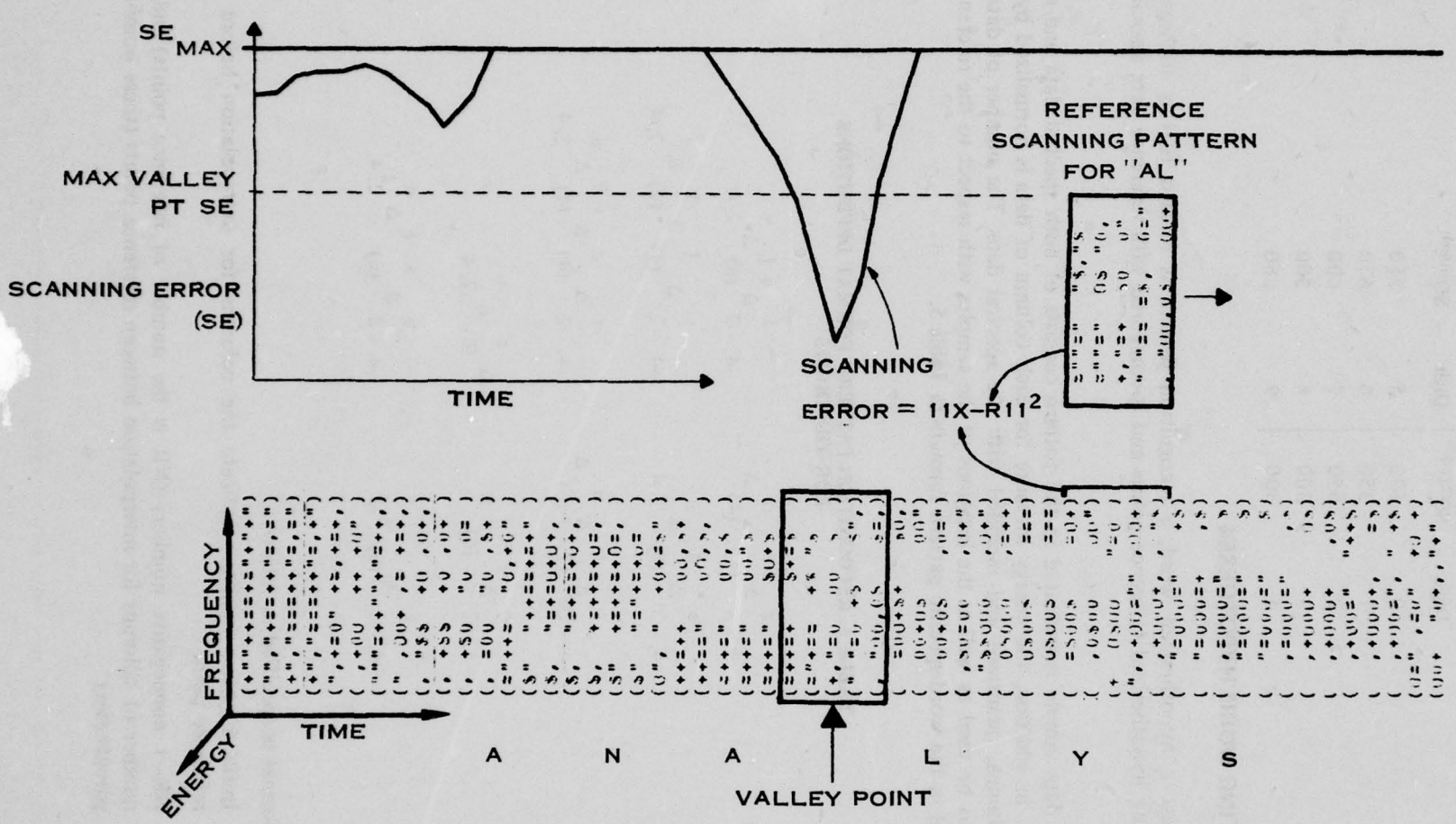


Figure 11. Example of Valley Finding

Digit	SQ^{thresh}	Digit	SQ^{thresh}
0	630	5	350
1	350	6	630
2	650	7	600
3	300	8	300
4	300	9	260

D. TESTING WORD HYPOTHESES

To test a hypothesized word, a recognition pattern is formed that is anchored at the corresponding hypothesized reference points and is compared with reference data associated with the word.

Each data sample represented in the pattern consists of both spectral data and regression coefficients. In addition, the energy measure for each column of data is normalized by the sum over all columns, quantized and included with the spectral data. The number of data samples (columns) to be used as well as the locations of the samples with respect to the reference points are specified in the word-specific pattern formats in Table 5.

TABLE 5. RECOGNITION PATTERN FORMAT DEFINITIONS FOR THE DIGITS

z Δ I r Δ o Δ −4, −2 (4) (4)	f Δ a I Δ v −4, −2 (6) 2, 4
w Δ Λ Δ n −4, −2 (6) 2, 4	s Δ I Δ k Δ s −4, −2, (4) (2) 2, 4
Δ t Δ u Δ −4, −2 (2) (6)	s Δ ϵ Δ v ∂ Δ n −4, −2 (4) (4) 2, 4
θ Δ r i Δ −4, −2 (6)	Δ e Δ t (6) 2, 4
f Δ o Δ r −4, −2 (6) 2	n Δ a I Δ n −4, −2 (6) 2, 4

The format is specified as follows:

1. Initial negative numbers indicate the columns for extrapolation before the first reference point
2. NR−1 nonnegative numbers (NR is the number of reference points) indicate the number of columns for interpolation between reference points (these numbers are in parentheses)

3. Remaining numbers indicate the columns for extrapolation after the last reference point.

Figure 12 shows a sample of the spectral data portion of a recognition pattern being extracted from an input speech spectrum.

After a recognition pattern is formated for the hypothesized digits, the squared Euclidian distance (TE) is calculated to all reference recognition patterns for that word and the minimum distance is retained for use in calculating a total normalized error (NE) for the digit (k) given below.

$$NE_k = \frac{TE_k / \#\text{ of columns in k}}{\textit{normalizing constant for digit } k} + w_k \frac{SQ_k}{10\,(NR - 1)}$$

where SQ_k is the sequence error (previous subsection) for digit k.

The total normalized error (NE) for each hypothesized digit is compared to a threshold, and the digit is discarded if the error is above the threshold. In addition, if the average energy across the word is less than a threshold (= 150), the digit is also discarded.

E. DIGIT SEQUENCE RECOGNITION

The digits remaining from the testing of the hypothesized digits are placed in a table. The final step then is to construct the six-digit sequence having the minimum error (ΣNE) for use in identifying the claimed identity for the speaker verification portion of the program. Since only 300 sequences were required in this study, certain constraints were imposed on the sequence to improve sequence recognition.

First, to reliably find reference points, certain digit combinations were disallowed as discussed in the section on segmentation. In addition to the types of transitions listed in that section that were disallowed due to lack of spectral transitionitivity, all nasal-to-vowel, glide, or semivowel (or vice-versa) transitions were disallowed because of the affect of nasals on the formants of adjacent vowels. Hence, the following digit combinations were not allowed:

0–1	2–8	3–9	7–1
0–8	2–9	4–1	7–8
0–9	3–1	4–8	9–1
1–8	3–8	4–9	9–8
2–1			

The second constraint imposed was that the first two digits in the sequence were linear combinations of the last four digits. Specifically,

$$d_1 = \left\{ \sum_{k=3}^{6} [(k-2)\, d_k]_{\text{mod } 11} \right\}_{\text{mod } 11}$$

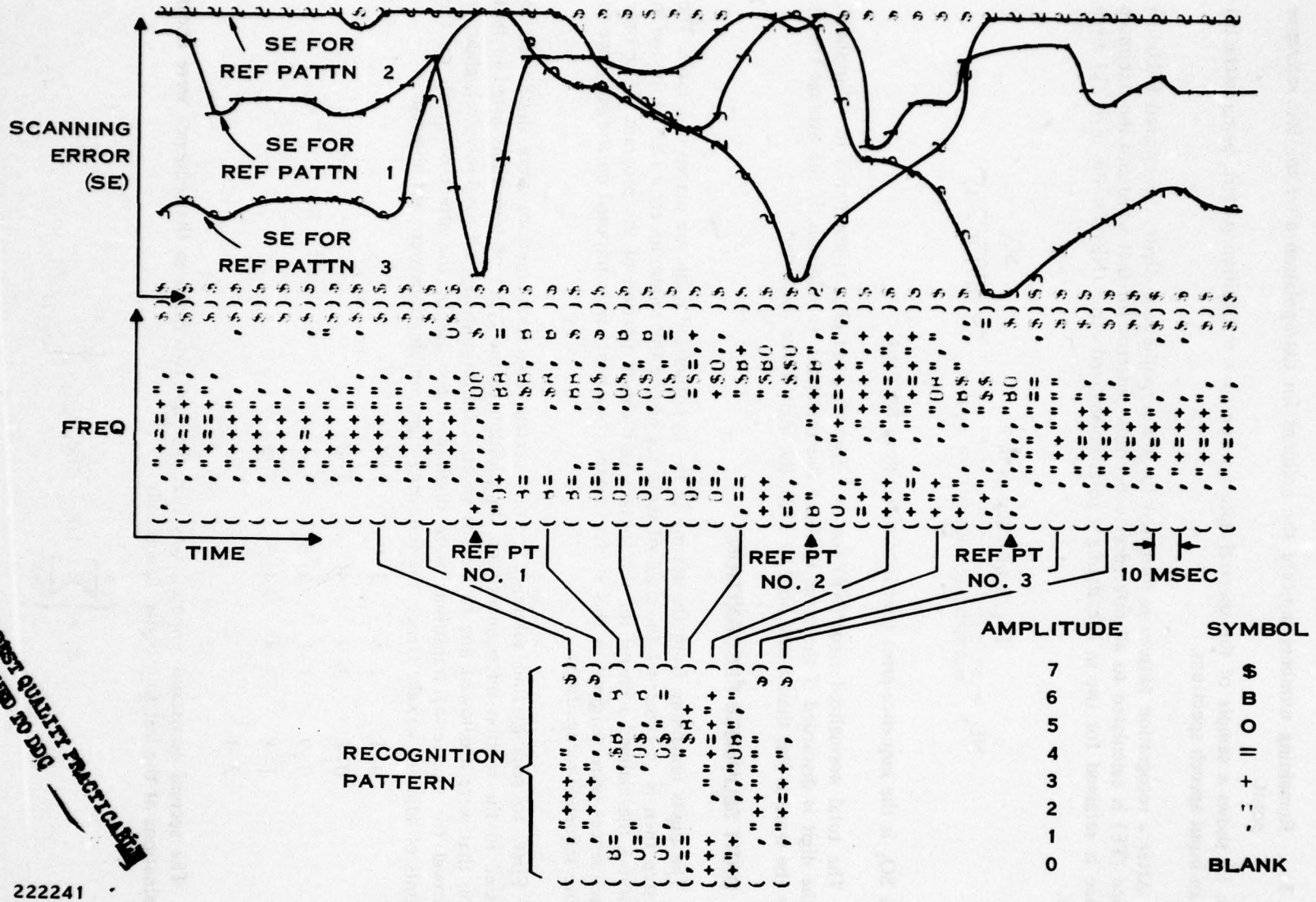


Figure 12. Locating Reference Points and Extracting Recognition Patterns

$$d_2 = \left[\sum_{k=3}^{6} d_k \right]_{\text{mod } 11}$$

for $d_k = \{0, 1, \ldots, 9\}$.

The third constraint was not for the purpose of aiding digit recognition, but rather to improve speaker verification performance. This constraint required that all six digits be different.

The resulting sequences after applying these three constraints are shown in Table 6.

The entries in the table of hypothesized digits are arranged in order of time of occurrence of the final reference point. A pointer exists for each entry to the location in the table of the first non-overlapping digit preceding it. The table is then searched for the six-digit sequence with the minimum sequence error, imposing constraints such as maximum interdigit times (0.3, 0.3, 0.6, 0.3, and 0.3 seconds, respectively) and maximum subsequence errors in order to minimize the search time. The specific algorithm is detailed in the software documentation for this contract.

After the minimum sequence error is found, it is compared to a threshold (= 480) to determine whether the sequence should be accepted or rejected, in which case the speaker would be prompted to repeat the sequence.

TABLE 6. ALLOWABLE SIX-DIGIT SEQUENCES

024587	026873	026954	027683	032651
035162	037459	045361	045876	047658
053274	054327	057261	057342	057423
061934	065174	068351	068432	068513
068947	072358	074563	076345	102479
104765	106547	107942	123687	125768
124063	124305	124659	125469	125973
126430	126945	127593	130275	132057
135072	137520	137954	142760	143570
145867	146758	147053	147508	150673
152374	153427	154237	158062	159034
159468	162573	162735	165327	167532
168342	168423	168504	173582	174635
175364	176840	179432	193042	193204
193476	194367	195702	195843	196572
196734	197463	197625	204675	204756
205647	206457	234687	234768	235810

TABLE 6. ALLOWABLE SIX-DIGIT SEQUENCES (Continued)

237945	243057	243561	246153	246587
255760	256190	258134	259610	260359
261754	265903	269305	273069	273654
275193	275436	275940	279504	305476
306952	307258	307681	307924	325874
342076	342580	345172	345687	345768
346810	347620	351627	356206	361079
361745	361907	365147	367514	368405
370468	403251	403685	406358	423579
426857	427051	430752	432615	435126
435207	437250	451960	455076	456172
457063	457306	458620	458973	461302
465275	465723	468072	468153	472583
476813	506934	510264	510426	512046
513794	516890	517943	520463	523640
540276	546207	547602	570369	576804
579234	581963	583079	592034	592468
594320	594673	596102	596374	597346
605872	612057	612703	613270	617934
619473	625170	625647	630572	635027
645307	645730	647350	647512	651942
653724	654372	658793	675823	675904
681520	681792	681954	683574	685194
685437	690243	690324	690758	693420
694230	694583	703658	705863	706592
706835	720364	723460	724351	724513
736251	740258	740681	743516	745136
745802	750519	751690	758946	759403
761052	762519	763590	764058	768045
790254	794302	802596	805269	805692
806745	810257	812604	812795	814062
814305	815034	815972	816359	817320
817592	819374	819536	820436	820517

TABLE 6. ALLOWABLE SIX-DIGIT SEQUENCES (Continued)

823451	824695	826043	827953	835270
835947	837052	840672	840753	845307
845127	847251	850367	851762	851924
852734	853706	854273	857612	861023
861457	869350	869512	873519	874590
875643	876453	876534	879045	879650
892430	893240	893402	895674	895103
896347	902567	905683	905764	906574
906817	907465	907546	920346	923604
925143	926034	927510	932750	935261
936152	937205	940582	942607	945037
945703	946270	946351	946513	951672
956127	957360	957603	958170	965372
965273	968027	968450	974581	975634

320 SEQUENCES OF 6 OBJECTS

SECTION III
THE DATA SET

Two data sets were collected for the digit sequence recognition study: a design data set and a test data set. The design data set consisted of recordings of one repetition of 1 of 10 possible sets of 10, six-digit sequences uttered by each of 560 subjects (333 males, 227 females), for a total of 5600 six-digit sequences (approximatly 6 hours of data). Data from 436 (230 males, 206 females) of the 560 subjects were collected over 5 days at Texas Instruments in Dallas, Texas. Very few of these subjects had any prior microphone experience. Data from the other 124 (103 males, 21 females) of the 560 subjects were collected over 3 days at Mitre Corporation, near Boston, Massachusetts. All but one of these 124 subjects were already participating in a speaker verification experiment being conducted at Mitre at that time,[12] and hence had microphone experience.

The test data set was collected for a speaker verification experiment at Texas Instruments in Dallas, and consisted of recordings of one repetition of one of ten possible sets of 10, six-digit sequences uttered by 106 subjects (64 males, 42 females). Each subject returned for several sessions (1 to 73; median no. = 25), usually on different days, over 3 months, using the same set of 10, six-digit sequences each time. Figure 13 shows the number of subjects participating in each session. In addition to this data, each subject had one (in some cases, two) "enrollment" session consisting of five repetitions of each of the 10, six-digit sequences in their assigned set. The total amount of data collected is between 60 and 70 hours.

The actual sequences used in both data collections are shown in Table 7. Typographical errors caused the sequences used in the Dallas portion of the design data set collection to differ slightly from those in the Boston portion, as noted in the table.

Both data sets are contained on annotated analog tape recordings. Background data were collected for most of the Dallas subjects (both data sets) including sex, birthplace, locations of all schooling, educational level, age, number of years in the Dallas area, parents nationality, and whether or not any speech defects existed. More limited background data had already been collected for the Mitre volunteers, consisting of sex, educational level, age, number of years in the Boston area, whether or not any speech defects existed and a general geographical region of primary schooling (there were four regions for the U.S., for example).

Data collected for both data sets in the Dallas area were recorded on a Teac 4010 SL, reel-to-reel tape recorder using an Electro-voice 674 microphone with a model 376 windscreen. Data collected in the Boston area used the same microphone, but were recorded on a more portable Sony TC-105 tape recorder.

NUMBER VERIFICATION SESSIONS	NUMBER OF PARTICIPANTS
1	2
2	4
3, 4	0
5	2
6	1
7–9	0
10	2
11	3
12	1
13	1
14	2
15	1
16	1
17–19	0
20	4
21	4
22	4
23	2
24	2
25	30
26	15
27	6
28	4
29	2
30	3
31	3
32	2
33–34	0
35	1
36	1
37	0
38	1
- - -	
46	1
- - -	
73	1

222242

Figure 13. Number of Sessions in Speaker Verification Data Base

TABLE 7. TEXTS USED IN DATA COLLECTIONS

Digit Sequence Recognition Texts

068947	068351	024587	032651	054327
159468	179432	168342	168423	168504
204675	204756	269305	279504	259610
305476	342076	342580	370468	306952
472583	476813	457306	427051	437250
513794	583079	581963	523640	512046
690243	681954	690324	*681792 (681972)	690758
751690	745136	736251	706835	743516
896347	850367	823451	814305	861457
937205	*907546 (907456)	925143	905683	905764
074563	053274	037459	*068432 (068342)	057261
135072	195762	125973	175364	158062
258134	256190	260359	243561	273654
307258	307681	361079	307924	351627
432615	458620	468072	461302	465723
579234	592034	592468	594320	594673
681520	619473	683574	630572	612037
763590	761032	764058	762519	740681
840672	824695	819374	869350	819536
906817	940582	946270	946351	936152

Speaker Verification Texts

057342	072358	027683	068513	061934
124063	145867	176840	165327	159034
273069	237945	243057	261754	253760
358206	361907	361745	346810	368405
451960	458973	468153	457063	430752
546207	510264	510426	520463	517943
612703	613270	675904	654372	675823
720364	724513	724351	759403	794302
869512	879045	879630	853706	852734
945703	946513	932750	942607	926034
035162	026954	047658	057423	026873
152374	162573	162735	142760	132057
206457	265903	269305	234768	234687
347620	367514	345172	345687	345768
453076	463275	423579	403685	403251
540276	570369	516890	576804	547602
658793	694583	694230	619473	651942
759403	758946	740258	750619	790234
851762	869350	861023	851924	879630
974581	951672	968027	968450	958170

*Sequences in parentheses are illegal sequences caused by typographical error. Sequences in parentheses used in Dallas; other three used in Boston.

SECTION IV
REFERENCE PATTERN GENERATION

A. CLUSTERING METHOD

Although some scheme, such as using scanning and recognition patterns for time registering the input speech waveform, is mandatory for speaker-independent word recognition, the drawbacks of using only one "representative" pattern for each reference point and for each word becomes readily apparent. The variations due to context, dialect, idiolect, and actual physical characteristics of the speaker (length and shape of vocal tract, pitch, etc.) must be accommodated by allowing multiple scanning and recognition patterns in order to gain acceptable performance levels.

Since the number of representative patterns needed was not known *a priori*, a hierachical clustering procedure was used on scanning and recognition patterns extracted from the design data set to determine sets of candidate clusters. Hierarchical clustering methods use a similarity matrix, whose entries measure the pair-wise similarities between the clusters in the data set, in determining which two clusters should be joined or divided. Iteratively joining clusters from N clusters to 1 cluster is called agglomerative clustering, and, less common, iteratively splitting clusters from 1 cluster to N clusters is divisive clustering. The entries in the similarity matrix are derived from either a distance or a correlation and are one of three types of measures:

1. Linkage (compares actual data between pairs of clusters)
2. Centroid (compares means of pairs of clusters)
3. Error sum of squares of variance (measures dispersion in new clusters resulting from the combination of all pairs of clusters).

More detail on hierarchical clustering can be found in Anderberg,[13] Everitt,[14] and Duda and Hart.[15]

The method used in this study was an agglomerative method that combined the two clusters having the smallest average distance between the points in the two clusters, i.e., combine the i and j clusters which have the minimum

$$\frac{1}{n_i n_j} \sum_{\vec{x} \in \chi_i} \sum_{\vec{x} \in' \chi_j} d(\vec{x}, \vec{x}')$$

where

n_i = the number of $\vec{x}$'s in class χ_i,

n_j = the number of $\vec{x}'$'s in class χ_j, and in this case

$d(\vec{x},\vec{x}') = \|\vec{x} - \vec{x}'\|^2$

The second step used was to iteratively improve on the partitions from the heirarchical clustering by moving samples from one group to another if such a move improves the value of some criterion function. This step used the iterative optimization method in Duda and Hart[15] that minimized the sum-of-squared-error criterion J_e, written as

$$J_e = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{\vec{x} \in \chi_i} \|\vec{x} - \vec{m}_i\|^2$$

where

$$\vec{m}_i = \frac{1}{n_i} \sum_{\vec{x} \in \chi_i} \vec{x}$$

If a point $\hat{\vec{x}}$ is moved from class χ_i to class χ_j, the means $\vec{m}_i$ and $\vec{m}_j$ change to

$$\vec{m}_i^* = \vec{m}_i - \frac{\hat{\vec{x}} - \vec{m}_i}{n_i - 1} \text{ and } \vec{m}_j^* = \vec{m}_j + \frac{\hat{\vec{x}} - \vec{m}_j}{n_j + 1}$$

J_i decreases to

$$J_i^* = J_i - \frac{n_i}{n_i - 1} \|\hat{\vec{x}} - \vec{m}_i\|^2$$

and J_i increases to

$$J_j^* = J_j + \frac{n_j}{n_j + 1} \|\hat{\vec{x}} - \vec{m}_j\|^2$$

Clearly then, since the criterion is to minimize J_e, if

$$\frac{n_j}{n_j + 1} \|\hat{\vec{x}} - \vec{m}_j\|^2 < \frac{n_i}{n_i - 1} \|\hat{\vec{x}} - \vec{m}_i\|^2 \tag{1}$$

then $\hat{\vec{x}}$ should be transferred from class χ_i to class χ_j. Specifically, $\hat{\vec{x}}$ is moved to the class χ_j, having the smallest $(n_j/n_{j+1}) \|\hat{\vec{x}} - \vec{m}_i\|^2$.

An additional property of this selection for J_e is that a set of equally divided clusters is favored over a set containing both small and large clusters. This can be seen by considering $n_i \gg n_j$ in equation (1), which yields approximately,

$$\frac{n_j}{n_j + 1} \|\hat{\vec{x}} - \vec{m}_j\|^2 < \|\hat{\vec{x}} - \vec{m}_i\|^2$$

Thus for $n_j = 1$, the distance $\|\hat{\vec{x}} - \vec{m}_j\|^2$ need only be less than twice the distance $\|\hat{\vec{x}} - m_i\|^2$ to the old mean to be transferred to class χ_j.

A flow chart for the clustering program including both the hierarchical clustering and the iterative optimization is given in Figure 14.

Returning to the hierarchical clustering, the question still remains of choosing what is the "proper" number of clusters. Rather than make the decision based on the clusters defined during the hierarchical clustering, the cluster definitions for 2 through 10 classes were each used in an iterative optimization, yielding new cluster definitions for each set of clusters. One set of these iteratively optimized clusters was then chosen on the basis of

1. A subjective judgment as to when no new unique features are present in the average patterns for each of the clusters.
2. Minimum value of (J_e for N clusters − J_e for N+1 clusters)/(J_e for N clusters).
3. Value of J_e (data sets with large J_e favor using more clusters).

In actually implementing the hierarchical clustering portion of this program, the question arose as to whether to calculate the similarity matrix once and update it or to recalculate the entries each step as needed. The similarity matrix contains N(N + 1)/2 unique entries. For even 32,000 computer words of storage, this yields only N = 252. In addition, a sequential updating procedure would tend to accumulate errors; hence, double precision would be used to help preserve accuracy, further reducing N for a fixed available amount of storage. In order to allow larger values of N, it was decided to recalculate the entries of the similarity matrix each step as needed. This approach was even more attractive because of the existence of a "vector comparator," a special purpose device that calculates $\| \cdot \|^2$ and is attached to the direct memory access port of the computer being used. This device is also used in the real-time processing for digit recognition.

B. CLUSTERING RESULTS

The clustering methods described in the previous subsection were used on both reference scanning and reference recognition patterns that were extracted from a subset of the design data set described earlier. This subset consisted of the first and third digits from each of 10, six-digit sequences from 85 subjects (42 males, 43 females). Since every digit appeared once in both the first and third positions of every set of 10 sequences, this provided a sample of both a short (first position) and long (third position) version of each digit from each of the 85 speakers. Reference points were automatically marked using some previously defined reference patterns and were all reviewed by hand to ensure their proper location. Using these reference points, scanning and recognition patterns were formed from all acceptable digits.

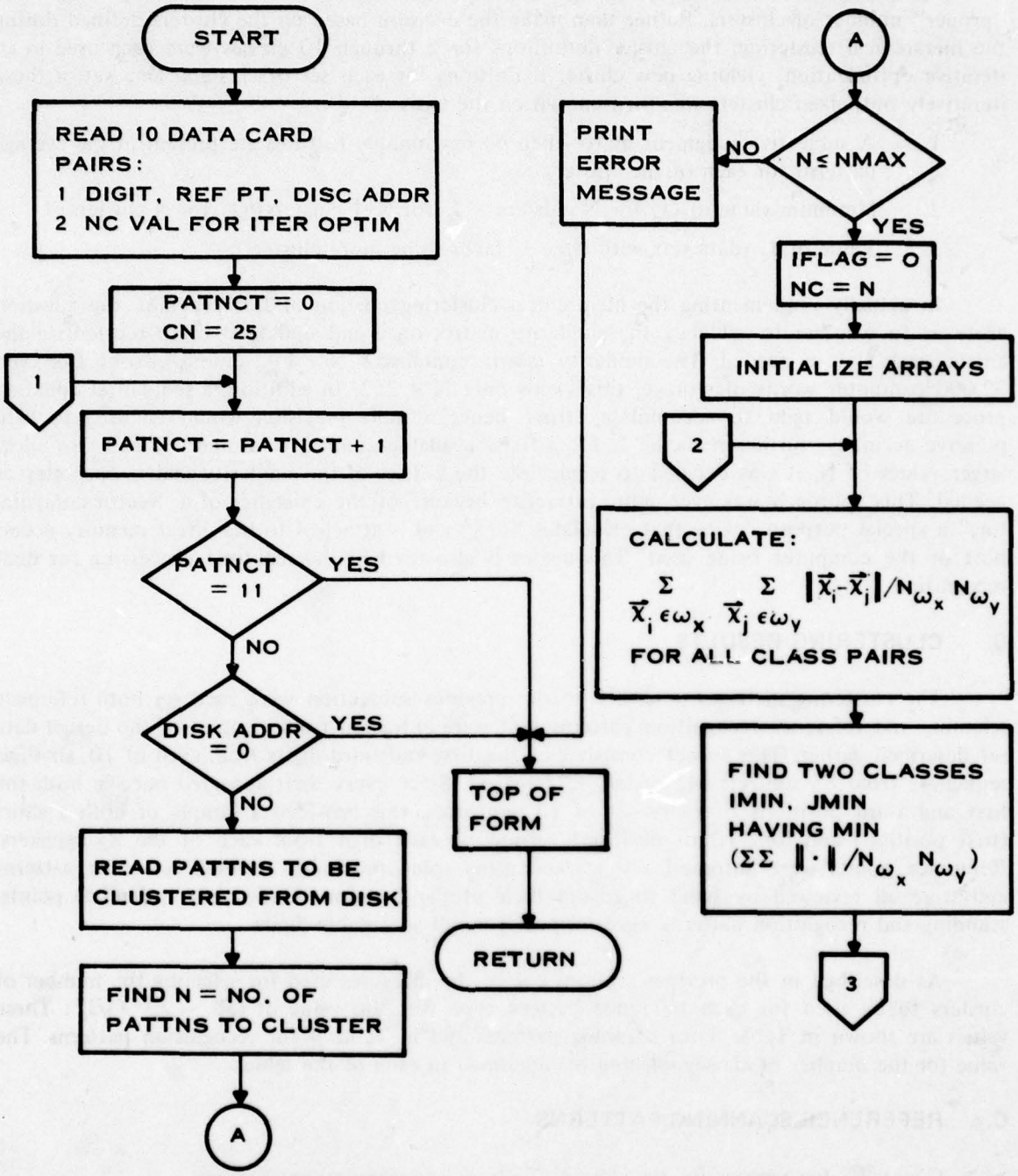
As described in the previous section, one of the measures used for selecting the number of clusters to be used for each reference pattern type was the value of $(J_e^N - J_e^{N+1})/J_e^N$. These values are shown in Table 8 for scanning patterns and in Table 9 for recognition patterns. The value for the number of classes selected is underlined in each of the tables.

C. REFERENCE SCANNING PATTERNS

Generally, the reasons for the class distinctions for scanning patterns are

Formant location differences (due to sex)

Lack of sibilant energy or aliasing of sibilant energy into lower filters (11–13) for females



222243

Figure 14. Flow Chart of Clustering Program (Sheet 1 of 5)

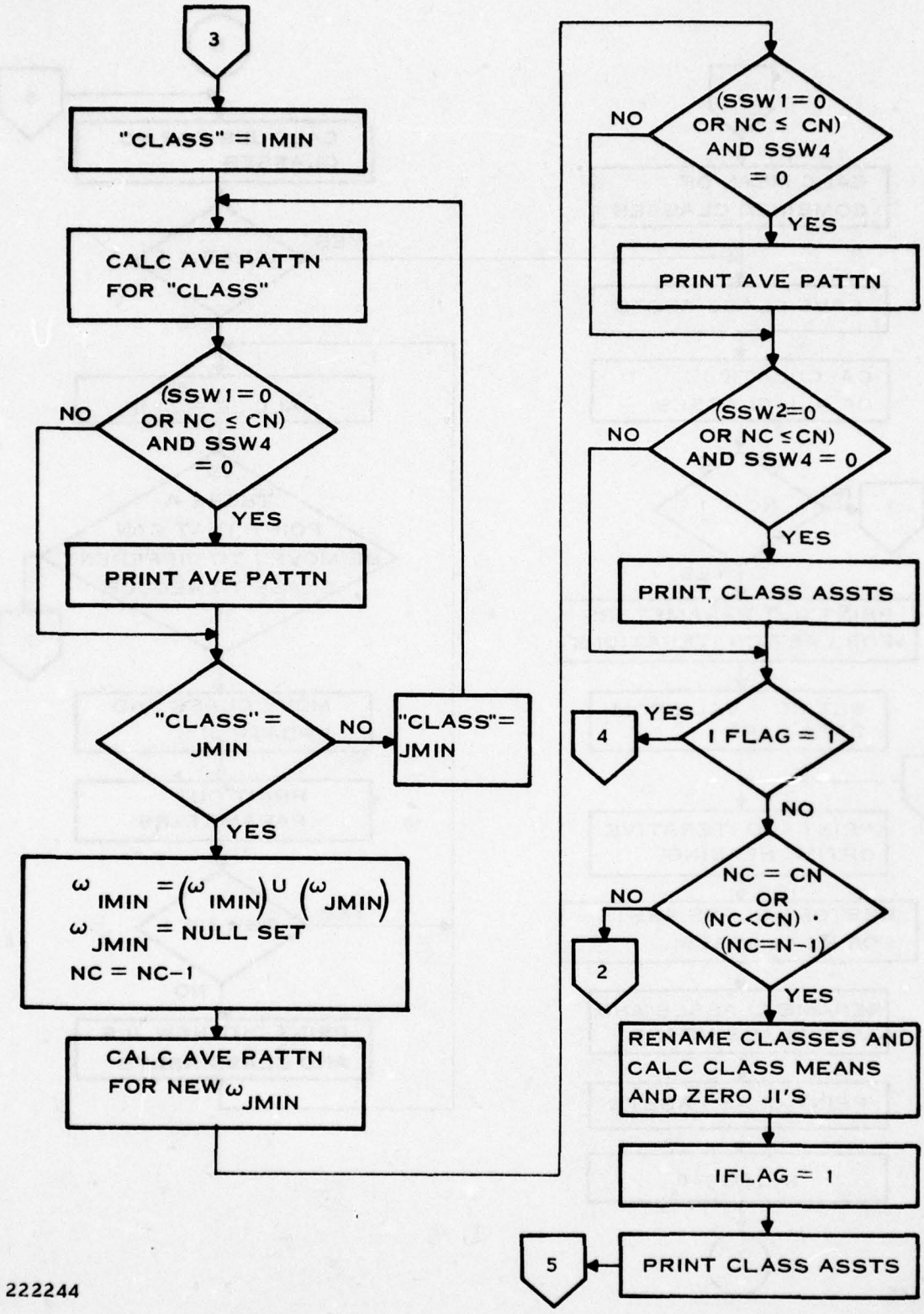


Figure 14. Flow Chart of Clustering Program (Sheet 2 of 5)

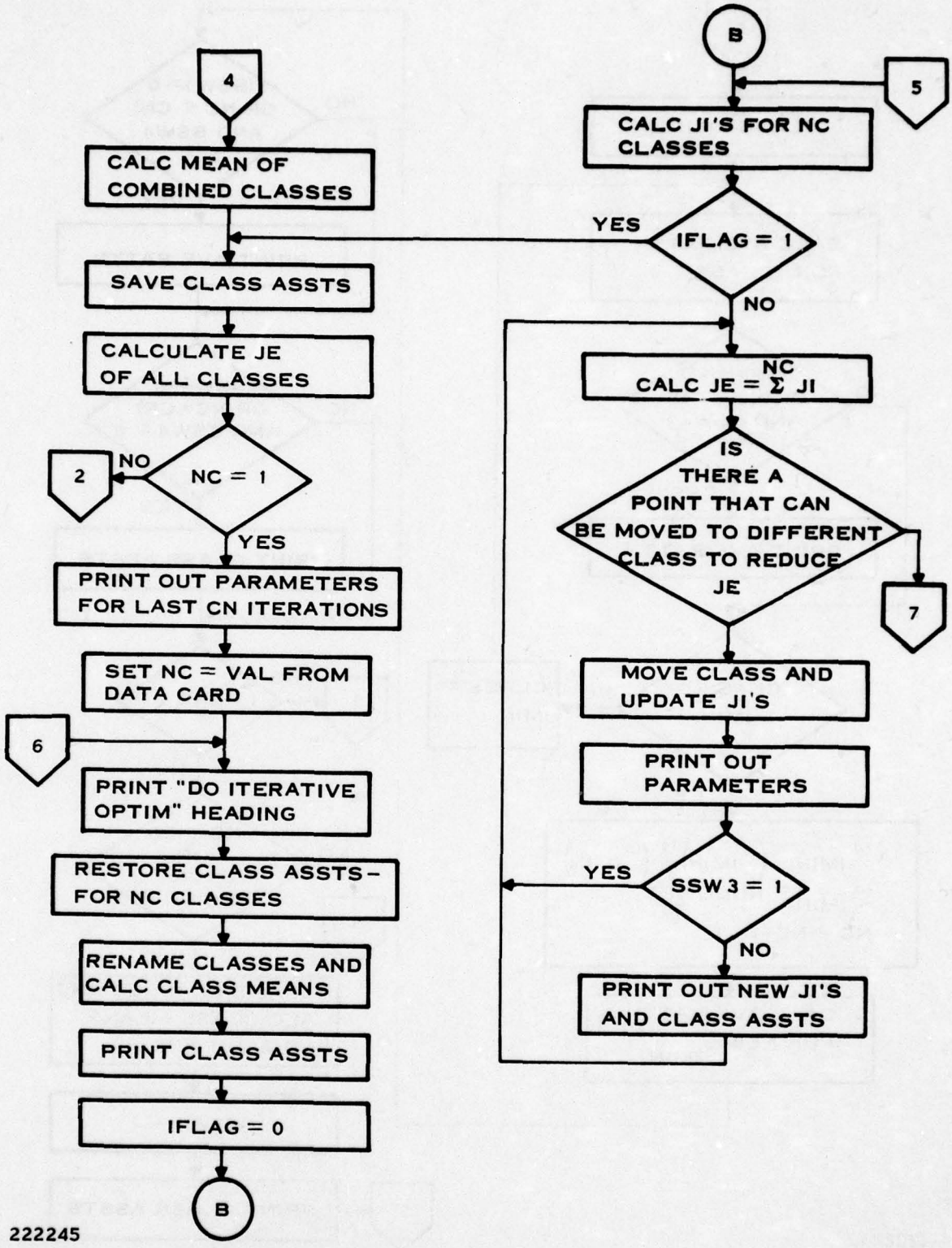


Figure 14. Flow Chart of Clustering Program (Sheet 3 of 5)

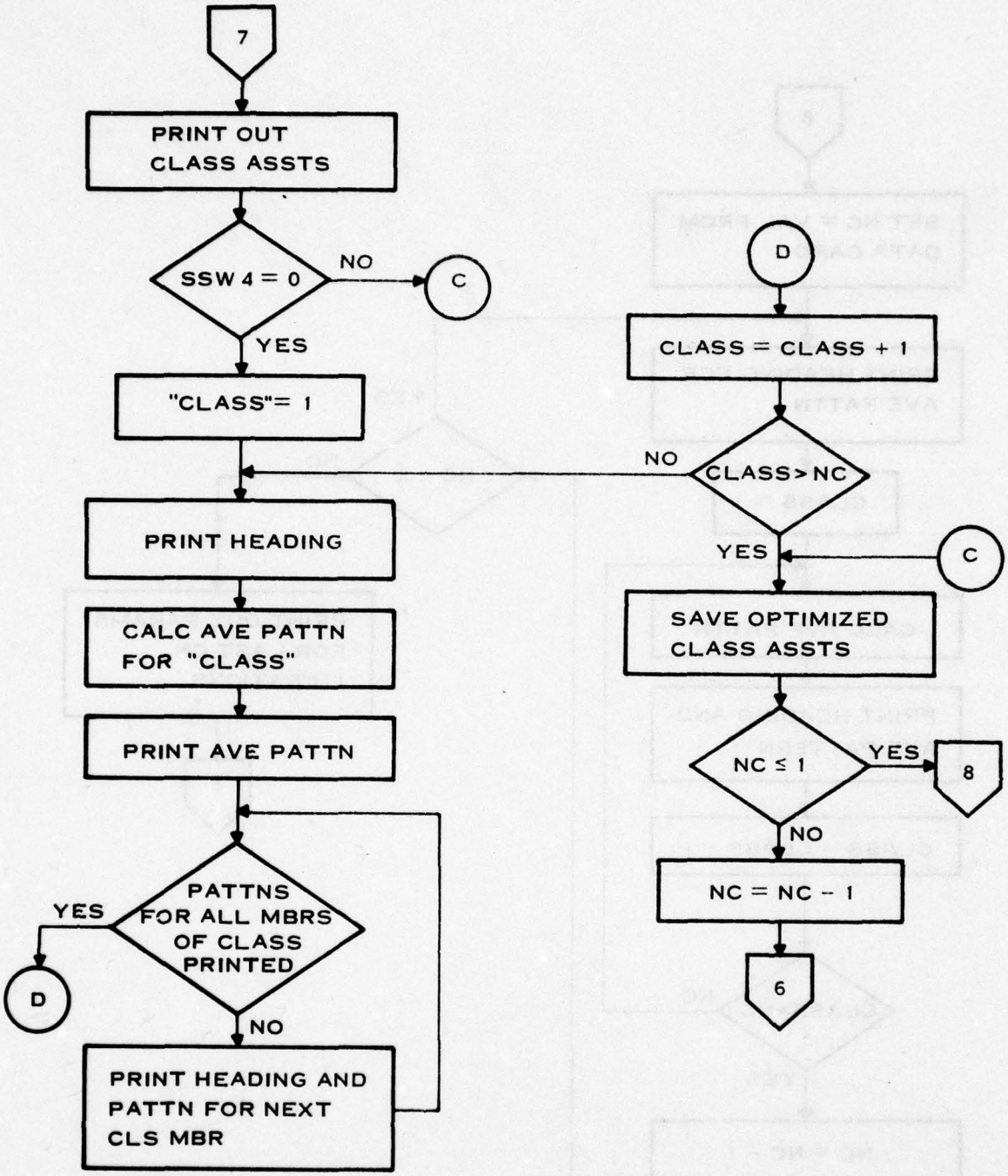


Figure 14. Flow Chart of Clustering Program (Sheet 4 of 5)

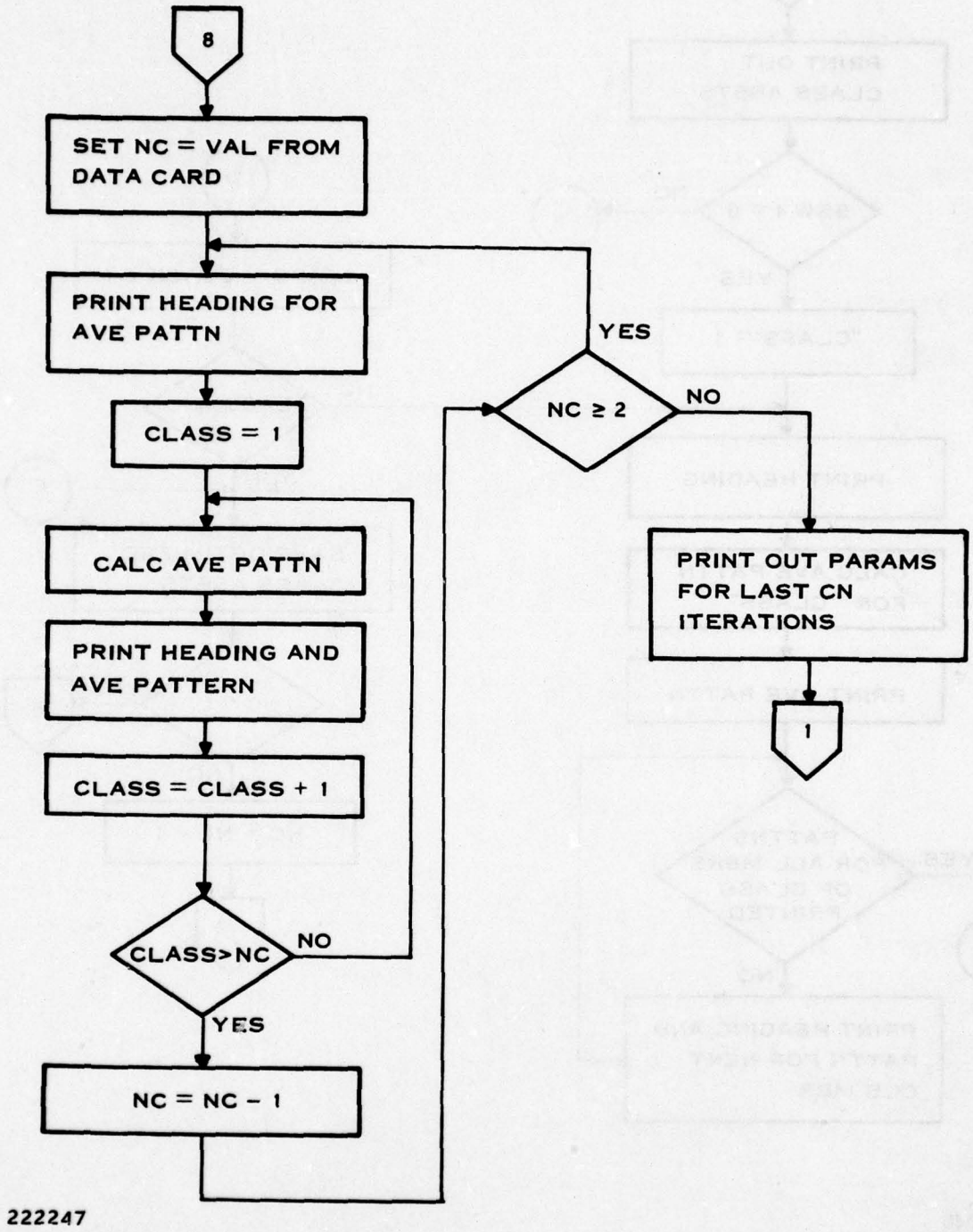


Figure 14. Flow Chart of Clustering Program (Sheet 5 of 5)

TABLE 8. SCANNING PATTERNS $(J_e^N - J_e^{N+1})/J_e^N$

No. of Classes	0			1		2			3		4	
	1	2	3	1	2	1	2	3	1	2	1	2
1	0.125	0.172	0.230	0.172	0.139	0.125	0.126	0.272	0.109	0.165	0.107	0.120
2	0.118	0.094	0.036	0.100	0.062	0.050	0.094	0.057	0.067	0.088	0.054	0.087
3	0.057	0.044	0.094	0.075	0.084	0.077	0.060	0.058	0.051	0.081	0.054	0.021
4	0.061	0.051	0.025	0.043	0.054	0.037	0.046	0.050	0.059	0.041	0.051	0.081
5	0.036	0.038	0.065	0.043	0.037	0.035	0.041	0.064	0.041	0.048	0.014	0.057
6	0.046	0.036	0.019	0.038	0.035	0.027	0.021	0.033	0.027	0.051	0.036	0.020
7	0.033	0.044	0.039	0.022	0.025	0.024	0.051	0.027	0.023	0.049	0.033	0.035
8	0.027	0.037	0.046	0.030	0.043	0.043	0.029	0.030	0.031	0.031	0.037	0.051
9	0.018	0.030	0.012	0.030	0.022	0.028	0.027	0.023	0.017	0.020	0.0004	0.024

No. of Classes	5		6			7			8		9	
	1	2	1	2	3	1	2	3	1	2	1	2
1	0.110	0.138	0.244	0.193	0.129	0.152	0.200	0.133	0.127	0.134	0.174	0.104
2	0.060	0.083	0.120	0.073	0.081	0.081	0.068	0.111	0.087	0.078	0.081	0.092
3	0.082	0.062	0.074	0.064	0.069	0.049	0.083	0.077	0.060	0.062	0.044	0.081
4	0.038	0.052	0.063	0.034	0.081	0.074	0.067	0.041	0.042	0.061	0.063	0.026
5	0.022	0.029	0.018	0.042	0.028	0.012	0.037	0.056	0.044	0.026	0.087	0.064
6	0.036	0.027	0.030	0.029	0.031	0.036	0.029	0.013	0.028	0.049	0.035	0.048
7	0.021	0.034	0.027	0.032	0.029	0.024	0.031	0.040	0.035	0.036	0.054	0.047
8	0.028	0.026	0.028	0.031	0.045	0.024	0.020	0.044	0.030	0.029	0.028	0.036
9	0.022	0.041	0.034	0.018	0.028	0.066	0.030	0.030	0.031	0.027	0.028	0.046

TABLE 9. RECOGNITION PATTERNS $(J_e^N - J_e^{N+1})/J_e^N$

	0	1	2	3	4	5	6	7	8	9
1	0.148	0.137	0.124	0.118	0.127	0.164	0.176	0.194	0.166	0.135
2	0.058	0.064	0.057	0.056	0.054	0.053	0.051	0.061	0.071	0.060
3	0.034	0.078	0.051	0.042	0.040	0.037	0.027	0.037	0.080	0.053
4	0.029	0.031	0.027	0.050	0.029	0.041	0.059	0.020	0.031	0.057
5	0.022	0.033	0.029	0.035	0.034	0.024	0.025	0.043	0.024	0.050
6	0.047	0.033	0.030	0.017	0.041	0.042	0.035	0.018	0.029	0.026
7	0.012	0.031	0.026	0.033	0.015	0.024	0.017	0.042	0.018	0.022
8	0.016	0.028	0.021	0.020	0.015	0.017	0.028	0.033	0.041	0.018
9	0.027	0.023	0.017	0.030	0.034	0.016	0.016	0.021	0.029	0.035

Vowel nasalization

Context differences

Presence or absence of the third format (F_3) for males during certain vowels.

A brief discussion of the pattern classes for each reference point for each digit follows. The figures showing each of the reference patterns have eliminated the difference data for purposes of brevity. The number of male and female contributors to each of the reference patterns is noted below each pattern.

Zero (Figure 15)

Ref. Pt. 1: /zI/

All patterns for this point contained a constant maximum amplitude for c_1 and a rising energy level from the sibilant into the vowel. Since /I/ and /i/ are close vowels in F_1/F_2 space, the pronunciation differences caused little pattern difference.

Male patterns: 1, 3, 6

Pattern 1 has no F_3 present during the vowel as patterns 3 and 6.

Pattern 6 has lower energy and a less prominent sibilant than the other two patterns.

Female patterns: 2, 4, 5

Pattern 2 has no sibilant present and a slightly higher F_1 than patterns 4 and 5.

Pattern 4, although having a more prominent sibilant than 2, has less energy during this part than 5.

Ref. Pt. 2: /ro/

This reference point presented the problem of choosing the point either when F_1 rose or when F_2 and F_3 split apart. Both events did not always occur, and when they did, they did not necessarily co-occur. Energy was relatively constant across all of these patterns.

Male patterns: 1, 5

Difference was in time of occurrence of F_1 movement

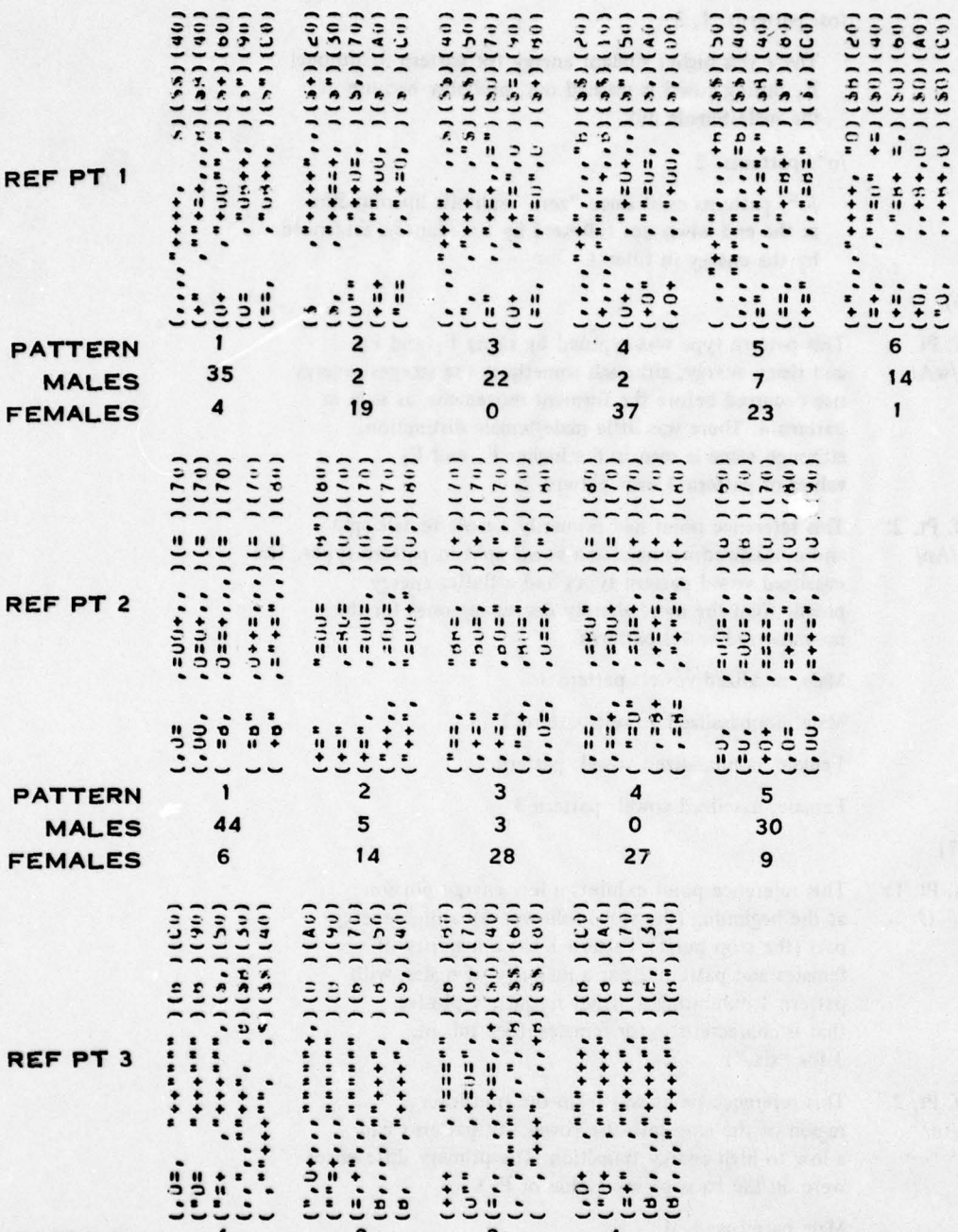
Female patterns: 2, 3, 4

Patterns 2 and 3 have no F_2/F_3 splitting

Pattern 3 has lower F_2/F_3 than patterns 2 and 4.

Ref. Pt. 3: /o^u/ /os/

For this reference point, the contextual split was more evident than that due to sex. Few of the vowel-to-sibilant patterns, however, were due to females due to the lack of prominent sibilants for females. In addition, the vowel formant structure appears washed-out, probably due to mixing of /o/ and /Λ/, whose second formants occur in different



222248

Figure 15. Scanning Patterns for "Zero"

filters. All patterns exhibited a decreasing energy profile.

/os/ patterns: 1, 3

There was higher sibilant energy for pattern 3, althougl F_1 during vowel is washed out, probably because of the male/female mix.

/o^u/ patterns: 2

/o^u/ patterns exist since "zero" exhibits lip rounding at the end when not followed by a sibilant as evidenced by the energy in filter 1.

One (Figure 16)

Ref. Pt. 1: /wʌ/ — This pattern type was typified by rising F_1 and F_2 and rising energy, although sometimes the steepest energy rise occurred before the formant movement, as seen in pattern 4. There was little male/female distinction, although some is seen in the higher F_1 and F_2 values of pattern 3 over pattern 2.

Ref. Pt. 2: /ʌn/ — This reference point had primarily a male/female split and a nasalized/nonnasalized vowel split in pattern types. The nasalized vowel pattern types had a flatter energy profile than the more sharply decreasing ones for the nonnasalized vowel patterns.

Male, nasalized vowel: pattern 1

Male, nonnasalized vowel: pattern 2

Female, nonnasalized vowel: pattern 3

Female, nasalized vowel: pattern 4

Two (Figure 17)

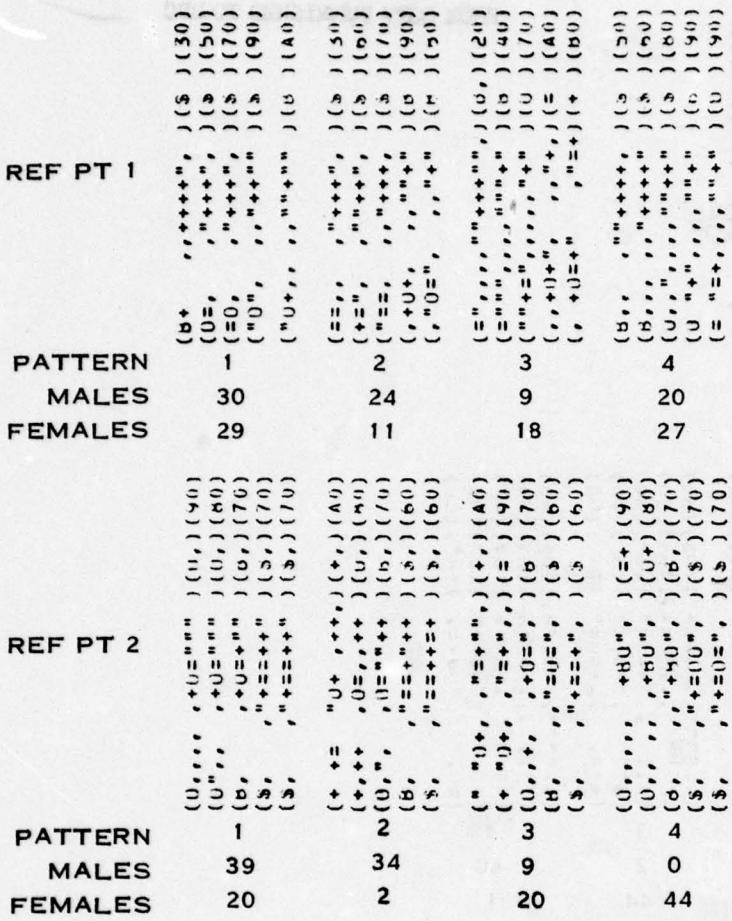
Ref. Pt. 1: /–t/ — This reference point exhibits a low energy portion at the beginning (the stop) followed by a high energy part (the stop burst). Pattern 1 has a majority of females and pattern 2 has a majority of males, with pattern 1 exhibiting a higher frequency plosive that is characteristic for females. (See ref. pt. 3 for "six.")

Ref. Pt. 2: /tu/ — This reference point was from the transition region of the stop into the vowel. All patterns had a low to high energy transition. The primary differences were on the basis of sex (value of F_2).

Male patterns: 1, 4

Female patterns: 2, 3

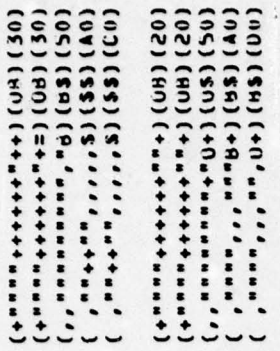
Ref. Pt. 3: /u–/ /us/ — The patterns for this reference point were split on the basis of context between /us/ as in the 2-6 or 2-7 context or /uz/ as in the 2-0 context, and /u–/ in the 2-silence context.



222249

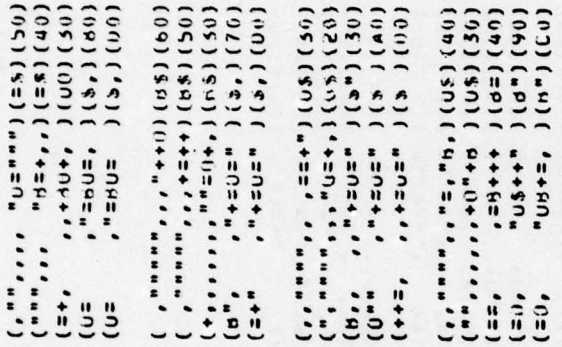
Figure 16. Scanning Patterns for "One"

REF PT 1



PATTERN	1	2
MALES	48	36
FEMALES	73	12

REF PT 2



PATTERN	1	2	3	4
MALES	36	6	2	40
FEMALES	9	31	44	1

REF PT 3

PATTERN	1	2
MALES	41	43
FEMALES	52	33

222250

Figure 17. Scanning Patterns for "Two"

Three (Figure 18)

Ref. Pt. 1: /θr/

The author cannot account for all of the distinctions among patterns for this reference point; however, patterns 3 and 4 are certainly distinguished by their lack of a sharp increase in energy and by the rapid decrease in F_2. This is caused by the presence of the "rolled-r" or flap in patterns 3 and 4, whereas patterns 1, 2, and 5 represent a single dental-to-palatal movement of the tongue.

Ref. Pt. 2: /i–/ /is/

Patterns for this reference point are divided primarily by context, with the first two representing the /i–/ context and the last three, the /is/ or /iz/ context. All energy profiles are generally decreasing.

/i–/ patterns: 1, 2

Pattern 1 is a female pattern, having F_2 primarily in filter 13

Pattern 2 is a male pattern, having F_2 primarily in filters 11 and 12

/is/ or /iz/ patterns: 3, 4, 5

These patterns are all characterized by energy in the top filters. Pattern 5 shows a characteristic of sibilants for some females, and is caused by the sibilant starting above the top filter, but being aliased down due to the lack of sharp cutoff on the anti-aliasing low-pass filter.

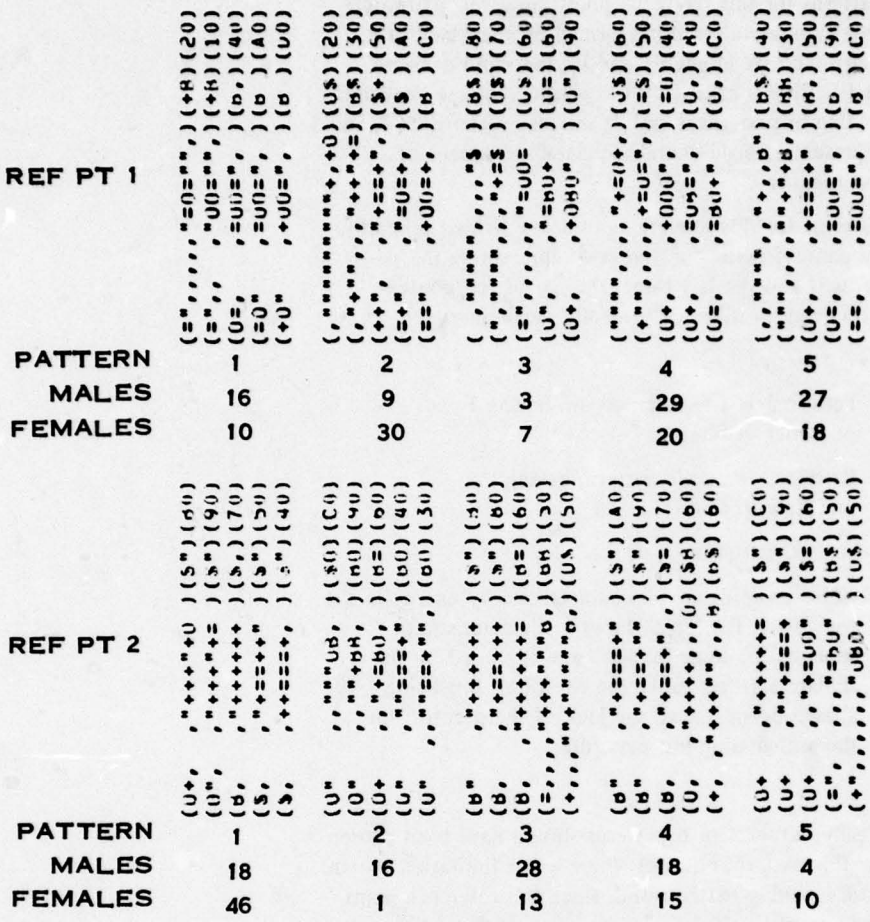
Four (Figure 19)

Ref. Pt. 1: /fo/

Ideally, either 4 or 5 patterns should have been chosen for this pattern; however, there was a limitation of 100 total scanning patterns and, since this reference point was one of the last processed, only one averaged pattern was used, resulting in very little sharp formant structure in the vowel. In the 4 and 5 pattern case, however, a distinction was made on the basis of sex due to the differing formant locations during the vowel. In addition, the 4 and 5 pattern cases showed higher relative amplitudes in the top filters, which became more "smeared" when only one pattern was used.

Ref. Pt. 2: /or/

The characteristics of this reference point location are a fairly sharp decrease in energy, a decreasing F_1, and a rising F_2 merging with a falling F_3. The problem that exists in locating this reference point is that the falling F_1 does not always coincide with the merger of F_2 and F_3 depending largely on the amount of lip rounding between the /o/ and the /r/; i.e., how much /u/, or even /w/, is inserted. The patterns split, however, largely on the basis of sex.



222251

Figure 18. Scanning Patterns for "Three"

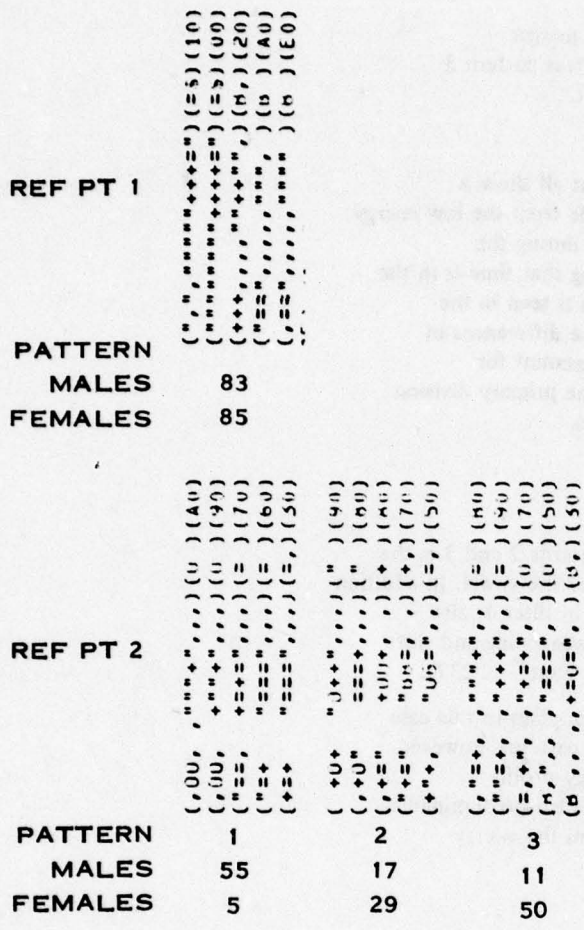


Figure 19. Scanning Patterns for "Four"

Male pattern: 1

Female patterns: 2, 3

Pattern 2 shows the F_2, F_3 merger preceding the F_1 drop, whereas pattern 3 shows them more coincident.

Five (Figure 20)

Ref. Pt. 1: /fa/

Patterns for this reference point all show a sharply increasing energy profile from the low energy /f/ to the vowel and a high c_2 during the fricative since the energy during that time is in the top filters. No dialect variation is seen in the reference patterns; however, the differences in F_2 between /a/ and /æ/ may account for the lack of a distinctive F_2. The primary division was between males and females.

Male pattern: 1

Female patterns: 2, 3

The distinction between patterns 2 and 3 is the appearance of nasalization of the vowel. In addition to the first nasal resonance in filter 1, also typical of nasalization is a weakening and shift up in frequency of F_1 (ref. Fant[17] p. 237).

Ref. Pt. 2: /Iv/ /ɛv/

The reasons for the pattern groupings in this case involves some speculation. All patterns, however, have a sharply decreasing energy profile (since the output of filter 1, which predominates during the /v/, is excluded from the energy calculation).

Male pattern: 2

This male pattern is for an /ɛ/ type of lower offglide (common in southern and southern mountain area speech), yielding the formants shown for this pattern.

Female pattern: 1

This female pattern is for an /ɛ/ type of lower offglide, yielding the formants shown for this pattern.

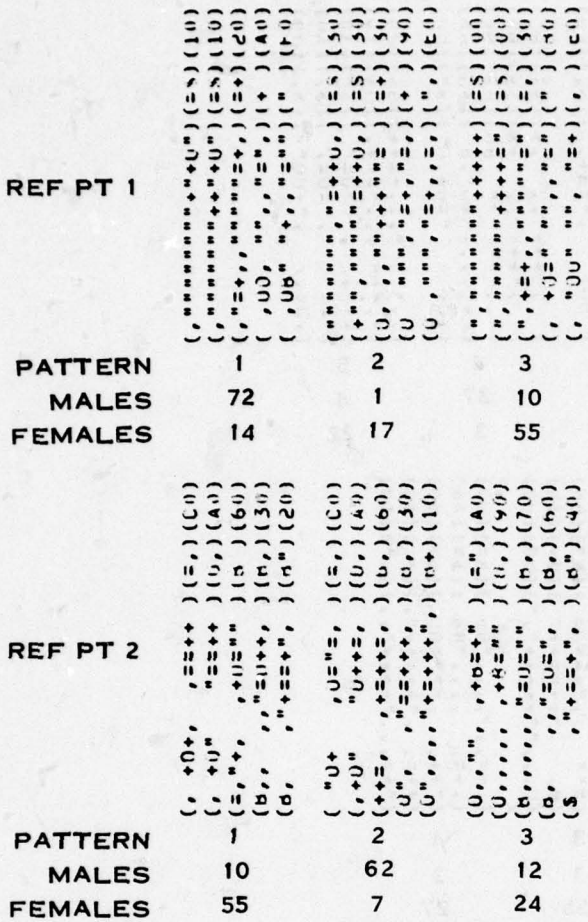
Mixture pattern: 3

It is proposed that the low F_1 is due to this pattern representing the high offglides /I/ and /i/ for both males and females.

Six (Figure 21)

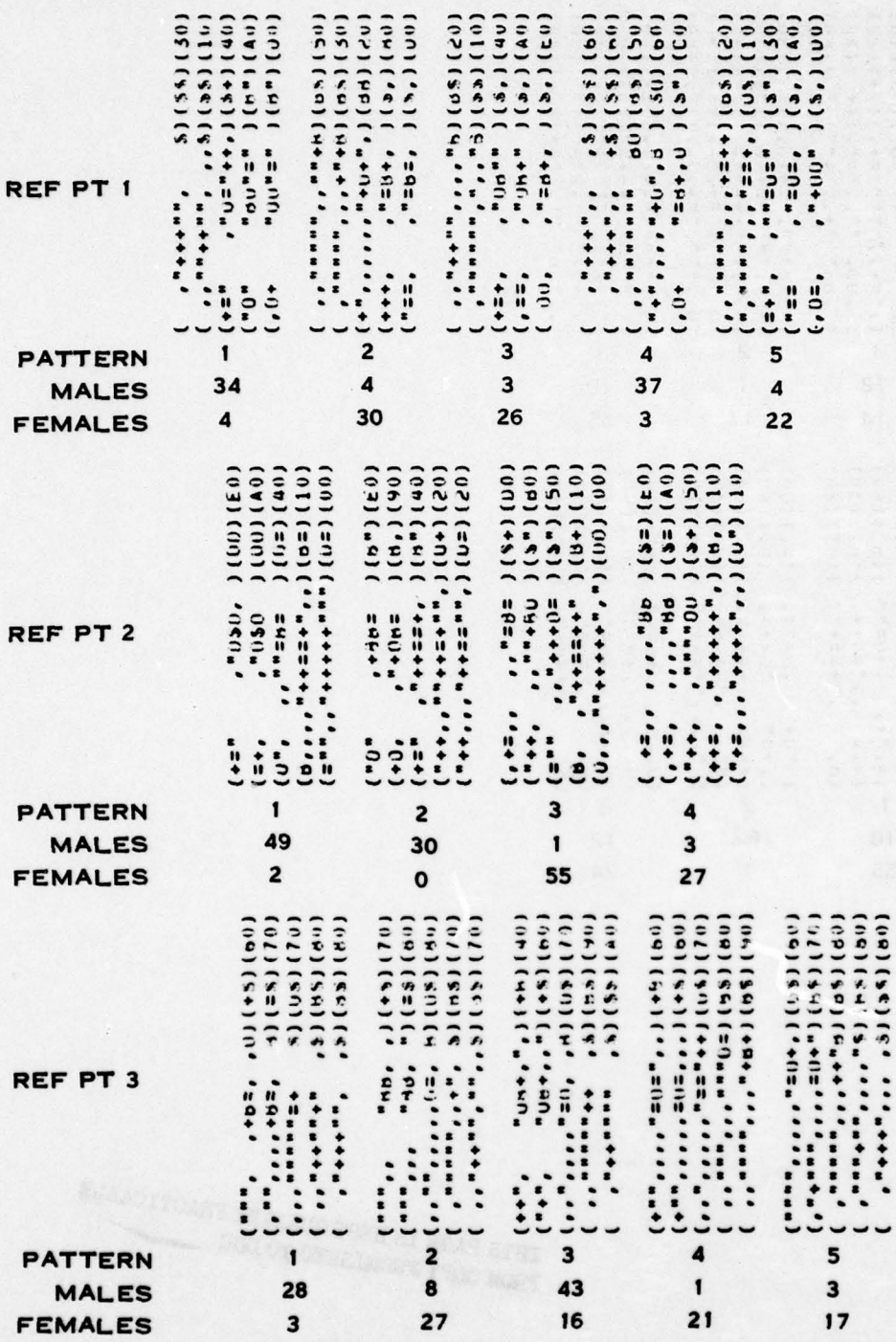
Ref. Pt. 1: /sI/

The classifications for this reference point are clearly on a male/female basis, which is consistent with the formant locations for the vowels.



222253

Figure 20. Scanning Patterns for "Five"



222254

Figure 21. Scanning Patterns for "Six"

Male patterns: 1, 4

Pattern 4 has a higher energy sililant and slightly higher, more prominent F_3 than pattern 1.

Female patterns: 2, 3, 5

The distinction between patterns 2 and 3 is primarily the higher energy in the sibilant for pattern 2. Pattern 5, however, is a result of the sibilant energy occurring above the top filter, but still below the cutoff of the input low-pass filter that aliases that energy down into filters 11–13.

Ref. Pt. 2: /l–/ This pattern is characterized by a sharp energy drop from the vowel to the stop. This division, as for the prior reference point, was on the basis of sex.

Male patterns: 1, 2

About the only distinction is the appearance of some filter 1 energy for pattern 1, indicating more of a glottal stop.

Female patterns: 3, 4

Again, about the only distinction is the filter 1 energy for pattern 3, presumably a more glottal stop.

Ref. Pt. 3: /ks/ The distinctions for this reference point are due to the higher plosive frequencies for females as for ref. pt. 1 of "two."

Male patterns: 1, 3

Female patterns: 2, 4, 5

Again, pattern 4 is distinctive because of the filter-bank anomaly of aliasing high frequency sibilants down into filters 12 and 13.

Seven (Figure 22)

Ref. Pt. 1: /sɛ/ Although there is some dialectic difference in perception of the vowel at this reference point between /ɛ/ and /æ/, no differences can be discerned because of the closeness in $F_1/F_2/F_3$ space of these vowels. The male/female distinction here is very clear however.

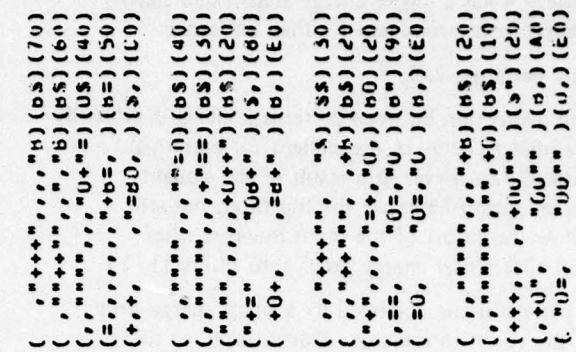
Male patterns: 1, 4

Two distinctions exist. One is the stronger F_3 for pattern 4; the other is the apparent earlier time registration of pattern 1.

Female patterns: 2, 3, 5

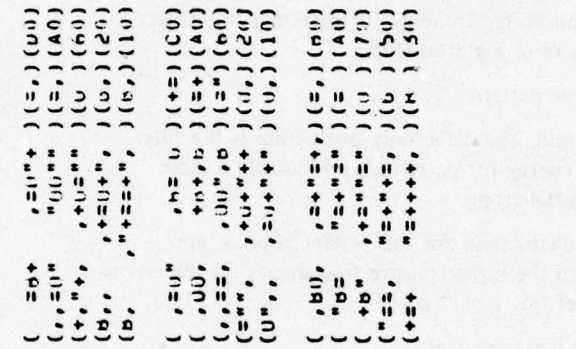
The differences among these patterns are the wash-out of F_1 for pattern 2 and the sililant aliasing in pattern 3, although not so striking as for ref. pt. 1 and 3 for "six."

REF PT 1



PATTERN	1	2	3	4	5
MALES	34	0	0	45	3
FEMALES	9	13	34	1	29

REF PT 2

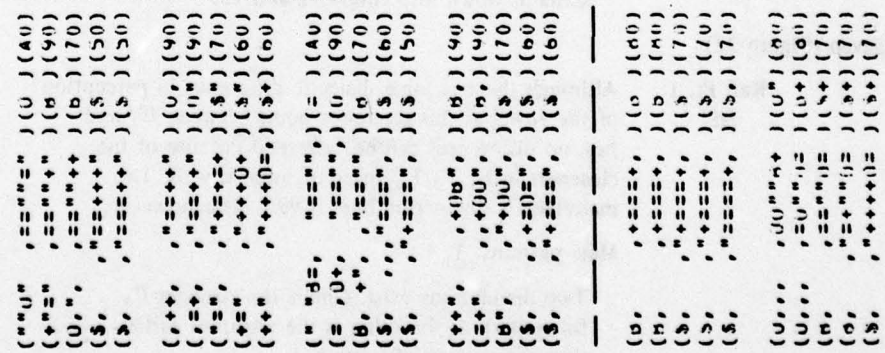


PATTERN	1	2	3	4
MALES	42	4	0	36
FEMALES	6	51	23	6

TYPE: NON-NASALIZED

TYPE: NASALIZED

REF PT 3



PATTERN	1	2	3	4	5	1	2
MALES	1	28	5	3	2	23	20
FEMALES	11	2	4	16	26	27	0

222255

Figure 22. Scanning Patterns for "Seven"

Ref. Pt. 2: /ɛv/

This reference point exhibits a sharply decreasing energy profile. The reference patterns were grouped on the basis of vowel formant frequency locations (i.e., sex) and on the basis of the amount of mouth closure for the labio-dental /v/ (i.e., the more closure, the lower F_1). Note also a reasonably strong F_2 during the labio-dental. Again, since /ɛ/ and /æ/ are close vowels, no dialect distinction was made.

Male patterns: 1, 4

The difference between 1 and 4 is the low F_1 during the /v/ in pattern 1. Since F_1 is in filter 1 (not included in the energy calculation), the energy for pattern 1 during the /v/ is lower than for pattern 4.

Female patterns: 2, 3

Although not as distinct as for the males, one difference between patterns 2 and 3 is the location of F_1 during the /v/. More of a distinction, however, is given by the F_3 during the vowel in pattern 3, which does not exist for pattern 2.

Ref. Pt. 3: /ən/

The major division between patterns for this reference point was between nasalized and nonnasalized vowels, the same distinction that was made for the "nine" scanning patterns. Whereas for all other patterns, the scanning error for each reference point was always chosen to be the minimum for each time sample, in the cases of patterns containing nasals in "seven" and "nine," two minimas are tracked: one for nasalized vowel patterns and one for nonnasalized vowel patterns.

Nonnasalized vowel:

Male pattern: 2

Female patterns: 1, 4, 5

Differences among patterns are presence of F_3 for pattern 1, and lack of F_1 and flatter energy profile for pattern 5.

Mixture pattern: 3

Although most nasals have an F_1 in filter 1, all do not, accounting for this pattern. However, the fairly equal mixture of males and females smears the formant structure of the vowels.

Nasalized vowel: Because of the energy in filter 1 during the vowel, note the flatter energy profiles than for the above patterns.

Male pattern: 2

Note the presence of the F_2 during the nasal in filter 12. Although this is supposed to be present during nasals according to Fant[17] only this pattern and pattern 3 of the nonnasalized vowel patterns show this.

Mixture pattern: 1

The strong F_1 in filter 1 during the vowel is present, but again, since this is a fairly equal mix of males and females the other formants during the vowel are smeared.

Eight (Figure 23)

Ref. Pt. 1: /–e/ — A sharply increasing energy profile is evident for all of these patterns.

/se/ — /se/ pattern: 1

This pattern represents the "six-eight" transition

/–e/ male pattern: 3

/–e/ female pattern: 4

Mixture pattern: 2

This pattern is a mix of /–e/ and /se/, both male and female patterns not close enough to patterns 1, 3, or 4 to be included with them.

Ref. Pt. 2: /et/ — All of these patterns have a sharply decreasing energy profile. In retrospect, too many patterns were chosen for this reference point. The two male patterns (1 and 3) are very similar and the only differences in the two female patterns (2 and 4) are in the higher F_1 in pattern 4 and a slight time registration difference. A mixture pattern (5) also exists, having a smeared F_2 due to the mixture of male and female F_2.

Nine (Figure 24)

Ref Pt. 1: /na/ — This reference point seems to show a distinctively regional flavor, since F_2 for all the vowels is in range for /æ/ rather than /a/. The nasalized, nonnasalized vowel distinction is made for both reference points of "nine", just as for ref. pt. 3 for "seven".

Nonnasalized vowel:

Male patterns: 2, 3

Pattern 2 has a higher C_2 during the vowel than pattern 3.

Female pattern: 1

Mixture pattern: 4

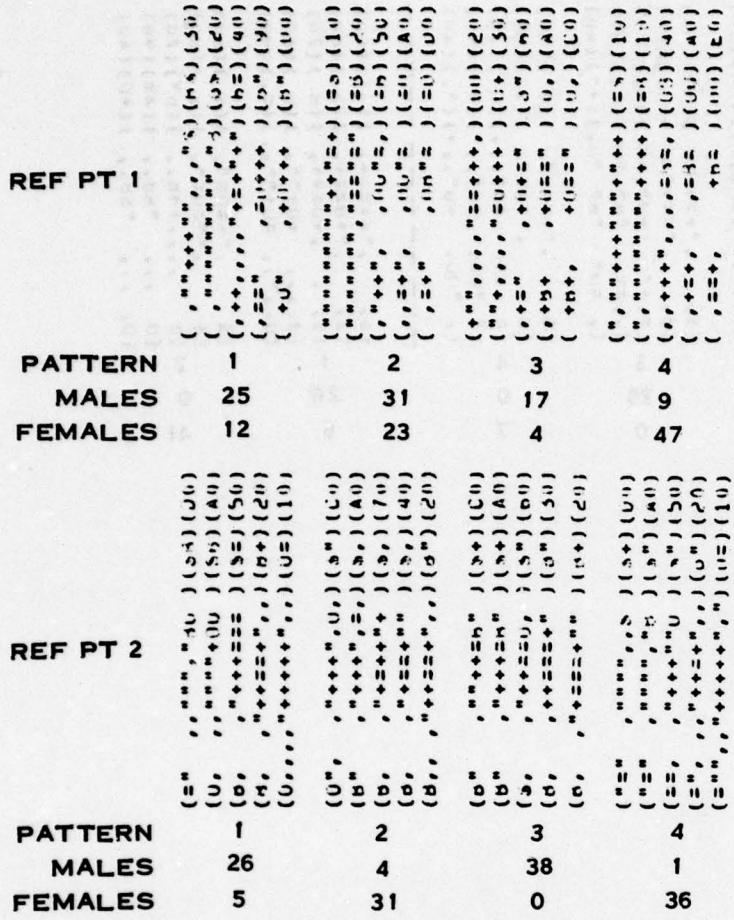
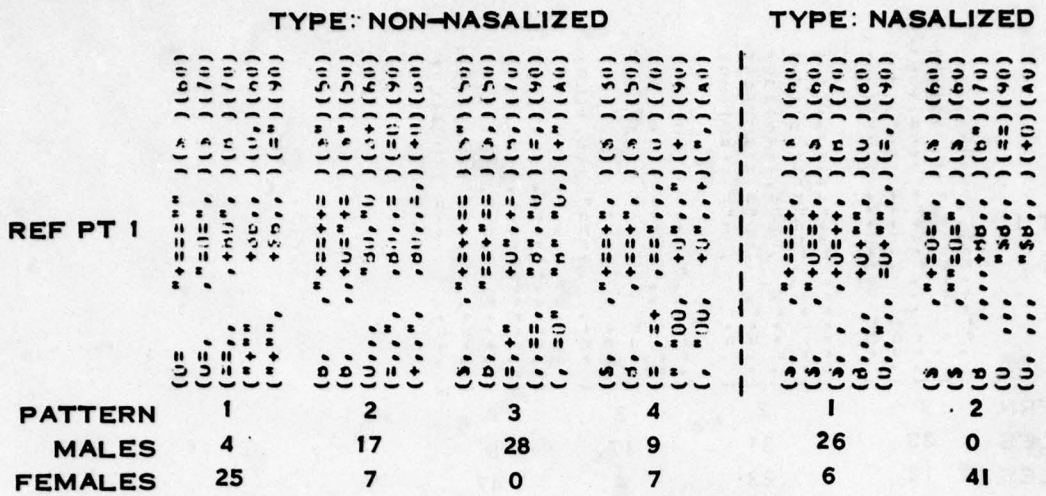
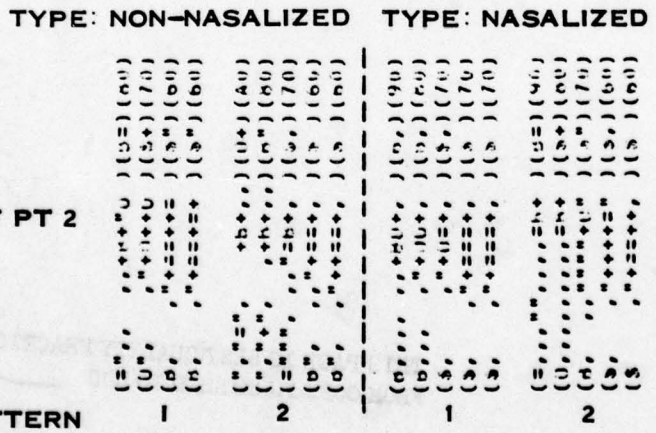


Figure 23. Scanning Patterns for "Eight"





222257

Figure 24. Scanning Patterns for "Nine"

Nasalized vowel:

Male pattern: 1

Female pattern: 2

Ref. Pt. 2: Again, the nasalized/non-nasalized distinction is made

Nonnasalized patterns: F_1 is not distinct for either of these two patterns, probably due to the mixture of offglides present (æ, ɛ, I), all of which have very close F_2 and F_3 locations, but have different F_1 locations.

Male pattern: 1

Female pattern: 2

Nasalized patterns:

Male pattern: 1

Female pattern: 2

D. REFERENCE RECOGNITION PATTERNS

Generally, the reasons for the class distinctions for recognition patterns are

Format location differences (due to sex)

Lack of sibilant energy or aliasing of sibilant energy into lower filters for females

Vowel nasalization

Dialect differences

Presence or absence of the third formant (F_3) for males during certain vowels.

A brief discussion of the pattern classes for each of the digits follows. The number of male and female contributors to each of the reference patterns is noted below each pattern.

Zero (Figure 25) Male patterns 1, 2

/zIro/

The only distinctive difference between these two patterns is the formant locations for the first vowel. Pattern 1 is closer to /i/ and pattern 2 is more representative of /I/.

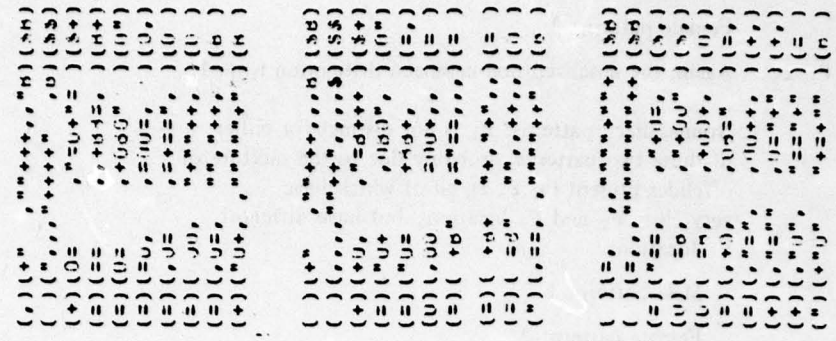
Female patterns: 3–7

Patterns 3 and 7 have missing sibilant energy, although pattern 3 still retained c_2 during the sibilant. Only pattern 5 seems to exhibit much lip rounding at the end as evidenced by the filter 1 energy.

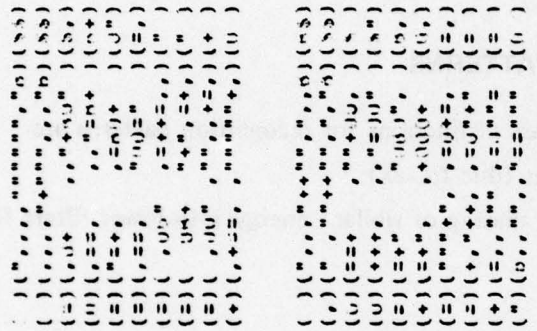
One (Figure 26) Male patterns: 1, 3

/wΛn/

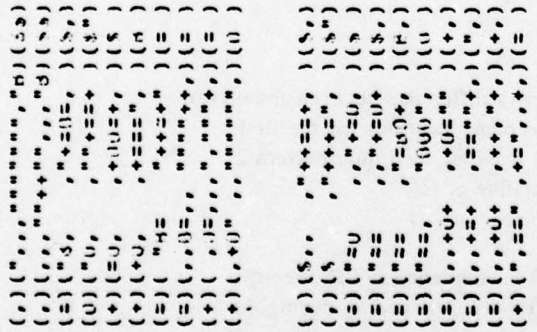
The difference between these two patterns is the vowel nasalization in pattern 3 and



PATTERN	1	2	3
MALES	41	36	1
FEMALES	0	2	16

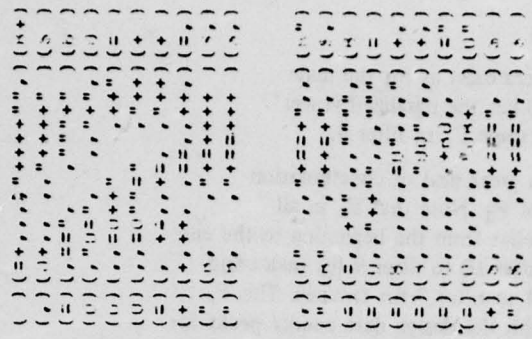


PATTERN	4	5
MALES	0	0
FEMALES	27	10

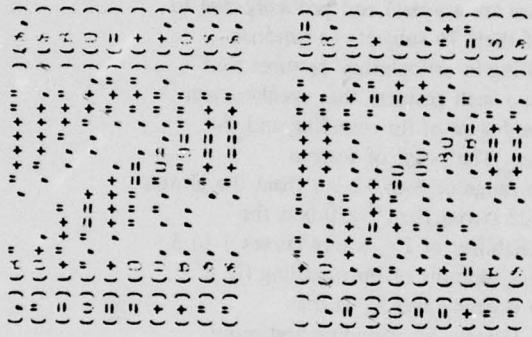


PATTERN	6	7
MALES	4	0
FEMALES	17	12

Figure 25. Recognition Patterns for "Zero"



	1	2
PATTERN		
MALES	46	1
FEMALES	5	28



	3	4
PATTERN		
MALES	31	4
FEMALES	18	35

Figure 26. Recognition Patterns for "One"

the F_1 present for pattern 1 in filter 4 during the vowel.

Female patterns: 2, 4

The same differences exist as for the male patterns. Pattern 2 has the nasalized vowel and pattern 4 has some F_1 in filter 4.

Two (Figure 27): /tu/

This pattern caused a great deal of consternation due to the location of F_2. Note that F_2 in all patterns drops somewhat from the beginning to the end of the vowel, from filter 10 to filter 9 for males and from filter 10 and 11 to filter 9 for females. This F_2 range is consistent with the design data energy peaks for F_2 for /u/, used in determining quantization thresholds, shown in Table 3, Section II. Note in this table, however, the disagreement with the Peterson–Barney[9] location for F_2 for /u/. It is proposed that the Peterson-Barney values are atypical and probably due to excessive coaching of their 76 subjects to produce "good" vowels, resulting in articulatory gestures that were exaggerated. Two such gestures that speakers can easily control are the degree of lip rounding and the location of the tongue. The point of tongue constriction is in the range of 8 to 12 cm from the glottis. Referring to Figure 28 (from Fant[17]), this is the region of greatest variability of F_2, where curves 1 to 5 represent an increasing amount of lip rounding (l_1 is the length of the lip opening and A_1 is the cross-sectional area). Greater lip rounding and more retracted tongue constriction produce lower F_2.

Male Patterns: 1, 2

The primary differences between these two patterns are in the location of F_1 at the beginning of the vowel and in the plosive energy and accompanying differences in c_2.

Female Patterns: 3, 4

The differences in these two patterns are due to more F_1 and F_2 movement in pattern 3, possibly indicative of more /Iu/ for the vowel, a dialectic variant more frequent in Texas.

Three (Figure 29): /θri/

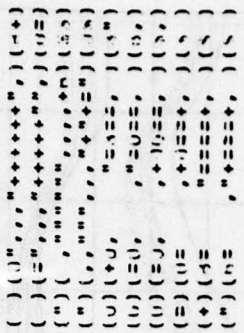
Male Patterns: 1, 3, 5

Pattern 3 shows the effect of a trill, or rolled /r/ by the dip in F_2. Pattern 5 has a higher F_2 due to the higher percentage of females in that cluster.

Female patterns: 2, 4, 6

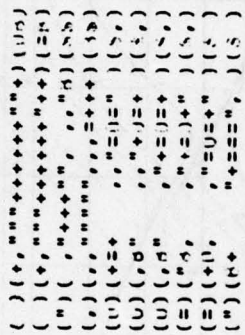
The differences in these patterns are minor. Pattern 2 does not have the c_2

PATTERN 1
MALES 55
FEMALES 8



PATTERN 2
MALES 29
FEMALES 7

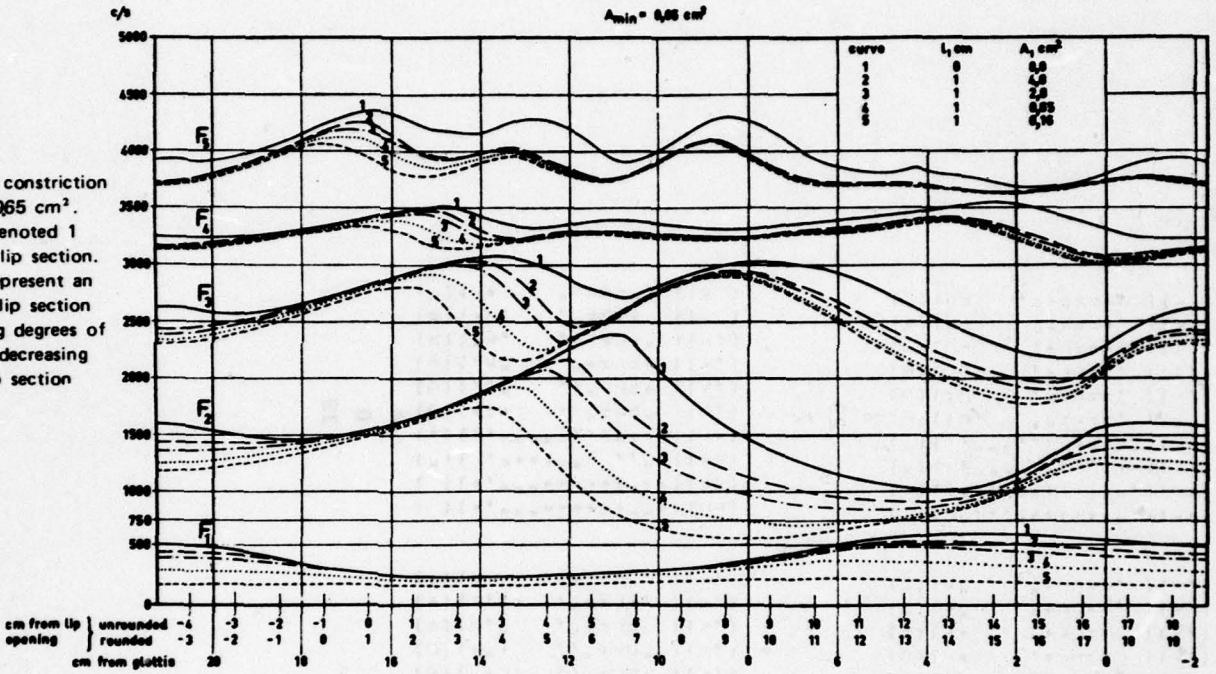
PATTERN 3
MALES 0
FEMALES 31



PATTERN 4
MALES 0
FEMALES 39

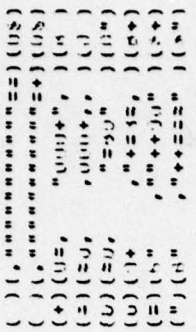
Figure 27. Recognition Patterns for "Two"

Note: Tongue constriction area A_{min} = 0,65 cm^2. Solid curves denoted 1 pertain to no lip section. Curves 2–5 represent an addition of a lip section with increasing degrees of rounding, i.e. decreasing area of the lip section (Fant, 1960)

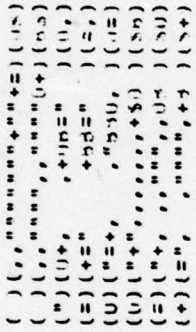


222261

Figure 28. Nomograms of the First Five Resonant Frequencies of the Vocal Tract Model



PATTERN	1	2	3
MALES	44	2	17
FEMALES	5	30	6



PATTERN	4	5	6
MALES	4	16	1
FEMALES	21	9	14

Figure 29. Recognition Patterns for "Three"

value during the /ri/ transition that the other two have. Pattern 4 has more of a steady state /r/ than the other two patterns.

Four (Figure 30) /for/

In retrospect, it is felt that too many reference patterns were chosen for the recognition pattern for "four," since these patterns do not show the sharp distinctions evidenced in other digits. Even though reasonably good separation was made on the basis of sex, the expected dialect variations did not appear.

Male Patterns: 1, 5

The small differences here are due to a slightly higher F_1 and more evidence of the rising F_2 in pattern 5.

Female Patterns: 3, 4, 6, 7

Noticeably lacking is any good formant structure except for the F_1 in patterns 6 and 7. The only other difference is in the c_2 value during the vowel.

Mixture pattern: 2

Missing in this pattern is the energy around filter 10 suggesting more of a /fow/ than /for/.

Five (Figure 31) /fa^{I}v/

Male patterns: 1, 5, 6

A dialect variation is very evident in these patterns, with patterns 1 and 5 having the diphthong and pattern 6 having only an /æ/. Patterns 1 and 5 are very similar with F_1 of pattern 5's vowel being higher.

Female patterns: 2, 3, 4

Pattern 2 in this grouping has the /æ/ without the diphthong. Pattern 3 differs from 4 by the filter 1 energy and the c_2 value during the vowel.

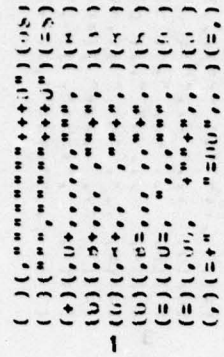
Six (Figure 32): /sIks/

Male Patterns: 1, 3

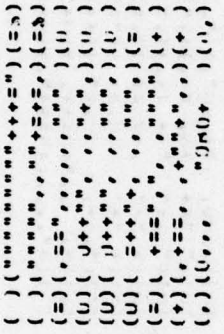
Pattern 3 exhibits a more positive spectral tilt than pattern 1 as shown by the stronger F_3 and c_2 and weaker F_1. Otherwise very similar.

Female patterns: 2, 4, 5

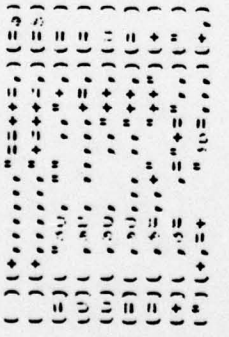
Primary differences in these patterns is due to sibilants. Pattern 2 has good initial and final sibilants; Pattern 4 has poor initial, but good final ones; Pattern 5 has poor sibilant representations in both positions. Again, this is a problem inherent



PATTERN	1	2	3
MALES	51	14	3
FEMALES	9	16	8



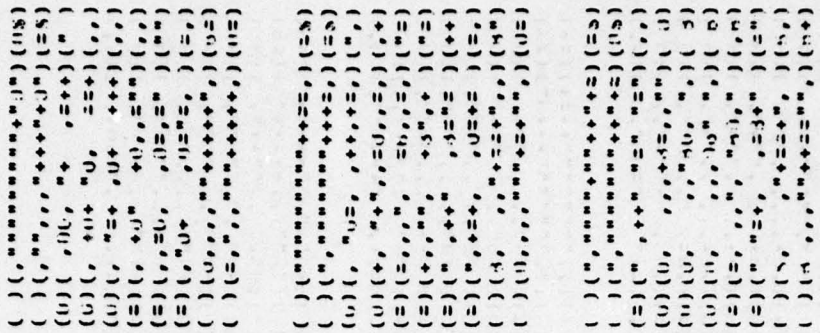
PATTERN	4	5
MALES	1	14
FEMALES	19	7



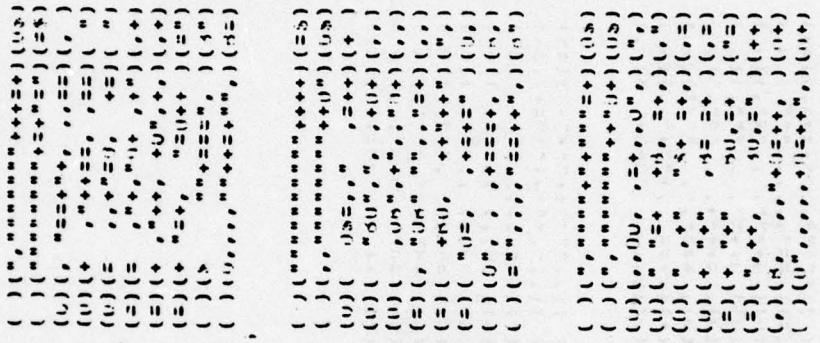
PATTERN	6	7
MALES	0	0
FEMALES	11	14

222263

Figure 30. Recognition Patterns for "Four"

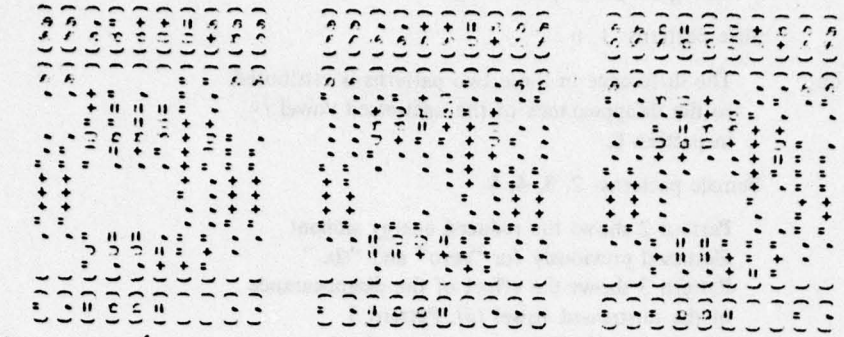


PATTERN	1	2	3
MALES	38	1	0
FEMALES	1	24	31

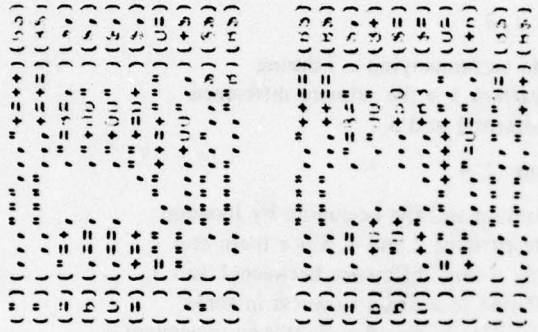


PATTERN	4	5	6
MALES	4	26	15
FEMALES	25	4	1

Figure 31. Recognition Patterns for "Five"



PATTERN	1	2	3
MALES	53	3	24
FEMALES	2	36	1



PATTERN	4	5
MALES	3	0
FEMALES	15	30

222265

Figure 32. Recognition Patterns for "Six"

in the filters due to the low sample rate. One final difference is the steadier F_2 during the /I/ in pattern 4.

Seven (Figure 33)

/sɛvən/

Male patterns: 1, 6

The difference in these two patterns is attributed to the disappearance of the unstressed vowel /ə/ in pattern 6.

Female patterns: 2, 3, 4, 5

Pattern 2 shows the reduced energy sibilant discussed previously for "zero" and "six." Pattern 3 shows the effect of the disappearance of the unstressed vowel /ə/. Pattern 5 has a strong third formant present (and accompanying c_2) during the /ɛ/.

Eight (Figure 34)

/et/

The formant movements for this digit show the actual pronunciation as /e^{I}t/.

Male patterns: 1, 3

The F_3 (and accompanying c_2) during the /e/ in pattern 3 is the primary difference between patterns 1 and 3.

Female patterns: 2, 4

Time registration and the beginning F_2 location differentiate patterns 2 and 4. Since there also appears to be a time difference between 1 and 3, this is attributed to a dialet slowness in some Texas speakers in starting the diphthong movement.

Nine (Figure 35)

/na^{I}n/

Male patterns: 3, 4, 5

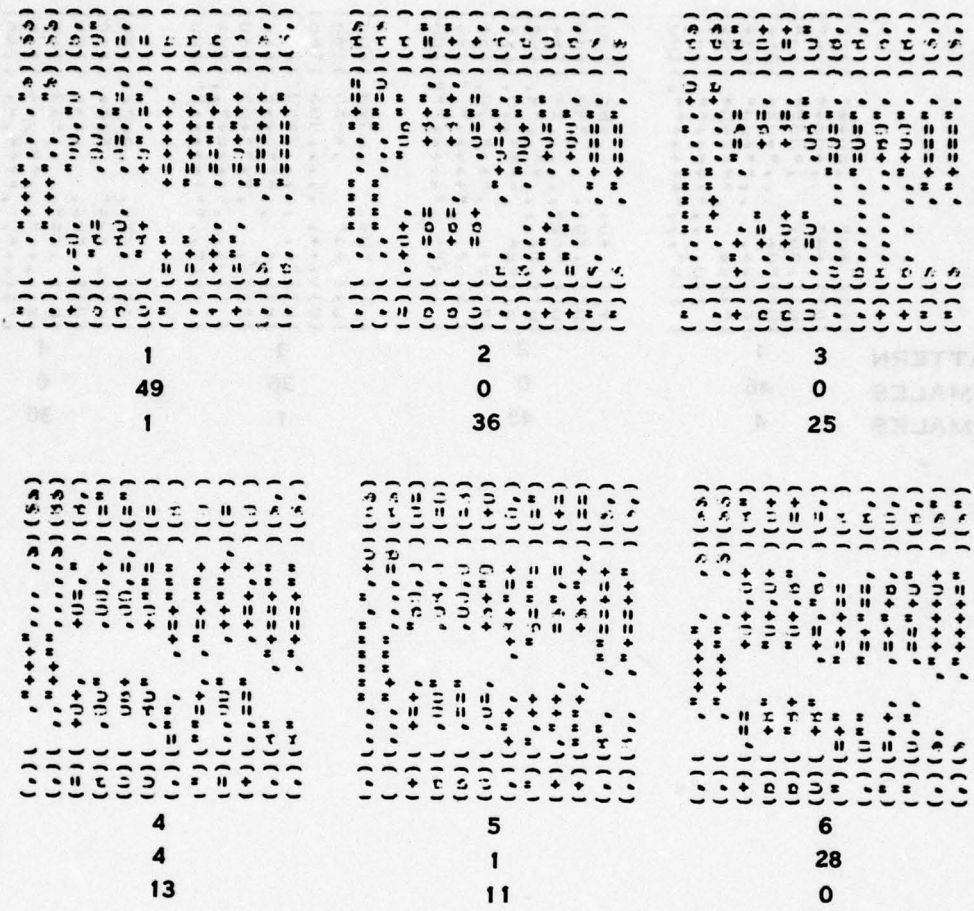
Dialect differences appear for these patterns with patterns 3 and 4 showing the diphthong rather than just the vowel /æ / shown in pattern 5. Additional differences are the F_3 in patterns 3 and 5 and the vowel nasalization (energy in filter 1) in pattern 4.

Female patterns: 1, 6

Dialect differences also appear for these patterns with pattern 1 showing the diphthong and pattern 6 the vowel /æ/. An additional difference is in the amount of nasalization during the vowel.

Mixture pattern: 2

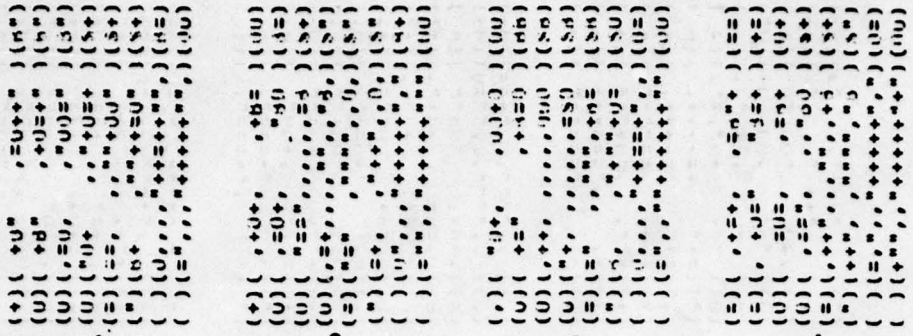
The feature distinguishing this pattern from male pattern 5 and female pattern 6 is the strength of F_1 and consequent lower value of c_2 during the vowel.



	1	2	3
PATTERN			
MALES	49	0	0
FEMALES	1	36	25

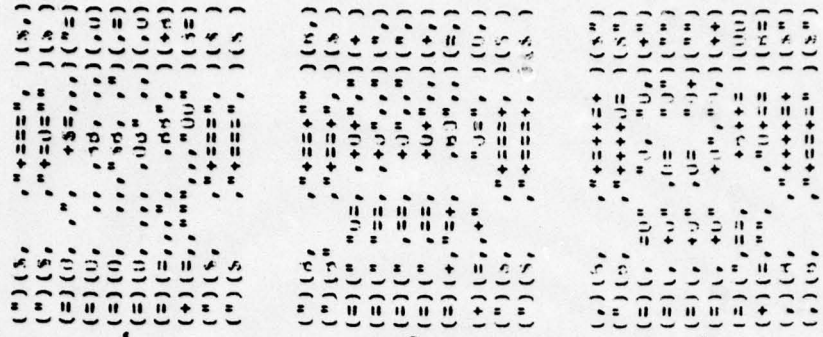
	4	5	6
PATTERN			
MALES	4	1	28
FEMALES	13	11	0

Figure 33. Recognition Patterns for "Seven"

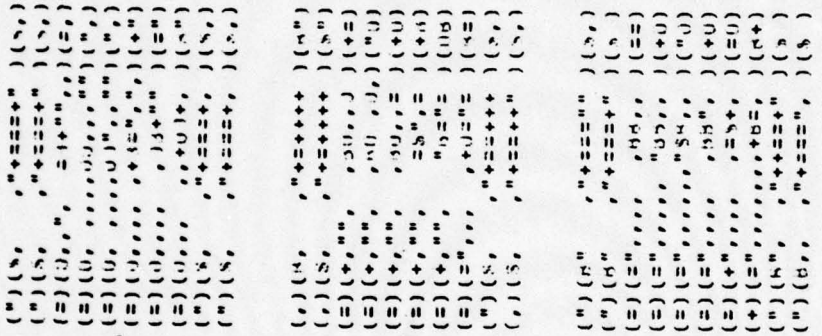


	1	2	3	4
PATTERN				
MALES	46	0	36	0
FEMALES	4	45	1	36

Figure 34. Recognition Patterns for "Eight"



	1	2	3
PATTERN			
MALES	0	13	18
FEMALES	39	10	0



	4	5	6
PATTERN			
MALES	26	27	0
FEMALES	5	0	32

222268

Figure 35. Recognition Patterns for "Nine"

SECTION V
SEQUENCE RECOGNITION TESTS

This section describes the sequence recognition testing done using a slightly modified version of the TVBISS program. The data base used for this program was one session of 10 six-digit sequences from each of the 106 speakers in the test data set described in Section III. These data were filtered, preprocessed, and stored in digital form on disk to provide a precisely replicatable data set. The evaluation program (TVEVAL) contained the same code as TVBISS except input data was from disk rather than being real-time data, no verification or enrollment was done, and the six-digit sequence found was compared to the known correct sequence. In addition, several types of data were saved on disk for each input sequence to allow subsequent compiling of digit recognition statistics and cumulative distribution plots for various parameters. An example of such a plot appears in Figure 40 showing the percentage of all digit "0"s having a total normalized error (NE) less than or equal to the value of the abscissa.

A. PARAMETER TESTING

The thresholds and parameters that are adjustable are enumerated below. The function of each of these parameters is explained in more detail in the section on speech processing.

- Reference point location parameters
  - Peak-to-valley ratio (PVR)
  - Maximum valley point error
- OPTSEQ (valley point sequencing) parameters
  - dt limits: dt_{max} , dt_{min}
  - Expected dt: $\hat{dt}$
  - Minimum expected dt (used to determine $\hat{dt}^*$ for the denominator in the point-pair error calculation): $\hat{dt}^* = \max(\hat{dt}, \hat{dt}_{min})$
  - Time deviation weighting = β
  - Floor of valley point error: OFFSET
- Hypothesized digit parameters
  - Minimum average energy across recognition pattern (EN_{min})
  - Weighting of sequence error (SQ) contribution to total normalized error for digit k: w_k
  - Normalizers to account for expected recognition error for digit k (TE_k)
- Maximum allowable total normalized error for digit k (NE_k)
- Six-digit sequence parameters
  - Syntactic constraints
  - Maximum interdigit times (time between first reference point of one word and last reference point of prior word)

Maximum subsequence error thresholds

Maximum sequence error.

A total of 45 evaluation runs were made using TVEVAL. The results and conditions for each of the evaluations runs from 5 on are given in Table 10. The starting baseline for run 5 is shown in Figure 36. The goals were to decrease running time, reduce the total error rate (rejection and substitution) to less than 2 percent, and to minimize the portion of error rate due to substitutions. Minimizing the portion of error rate due to substitutions is effected by moving the percent rejected curve left relative to the percent substituted curve in Figure 36.

In addition to varying parameters and thresholds, the early evaluation runs were used to determine the value of using multiple reference patterns. Run 5 used single reference recognition patterns and multiple reference scanning patterns. Runs 6 and 7 split the reference recognition patterns into several patterns that were the cluster averages for each digit, rather than having only one pattern for each digit. This resulted in lowering the recognition error (TE) and, hence, the total normalized error (NE) for each digit and the total sequence error (ΣNE). This shifted the percent rejected versus total sequence error curve to the left, as shown by comparing Figure 36 and Figure 37. Note that the percent substitution curve also moved to the left, but far less than the percent rejected curve.

Runs 9 and 10 split the reference scanning patterns for vowel-nasal and nasal-vowel transitions in "seven" and "nine" into two separate groups. Since the scanning error (e_i) is the minimum of the distance between the input data and each of the reference scanning patterns for reference point i, only one e_i resulted previously. This change produced two e_is for each of these three transitions, one group for nasalized vowels and one group for non-nasalized vowels. Figure 38 shows the results for evaluation 10, which uses split scanning patterns for both of these digits. Since several substitutions were changed into acceptances (see Table 10), the percent substitution curve moved right and the percent rejected curve moved to the left.

No further changes in either reference scanning or recognition patterns were made after run 10.

1. Peak-to-Valley Ratio (PVR)

Evaluation runs 20 through 24 maintained all variables except the peak-to-valley ratio (PVR) fixed. The PVR (discussed more thoroughly on pages 20 through 31 of Speaker Verification I^3, where PVR = 2) is the ratio of the scanning error following a potential valley point to the error at that valley point necessary in order to establish the existence of the valley point. Some comments follow.

Lower PVRs produce more valley points, increasing processing time and decreasing the error rate for missed reference points. Specifically, some numbers generated from a 120-word data set said in an isolated manner from a single speaker follow:

PVR	No. of Valley Pts	Percent Missed
2.0	360,220	3.12
1.5	668,526	2.58
1.25	927,832	0.54
1.125	1,124,668	0

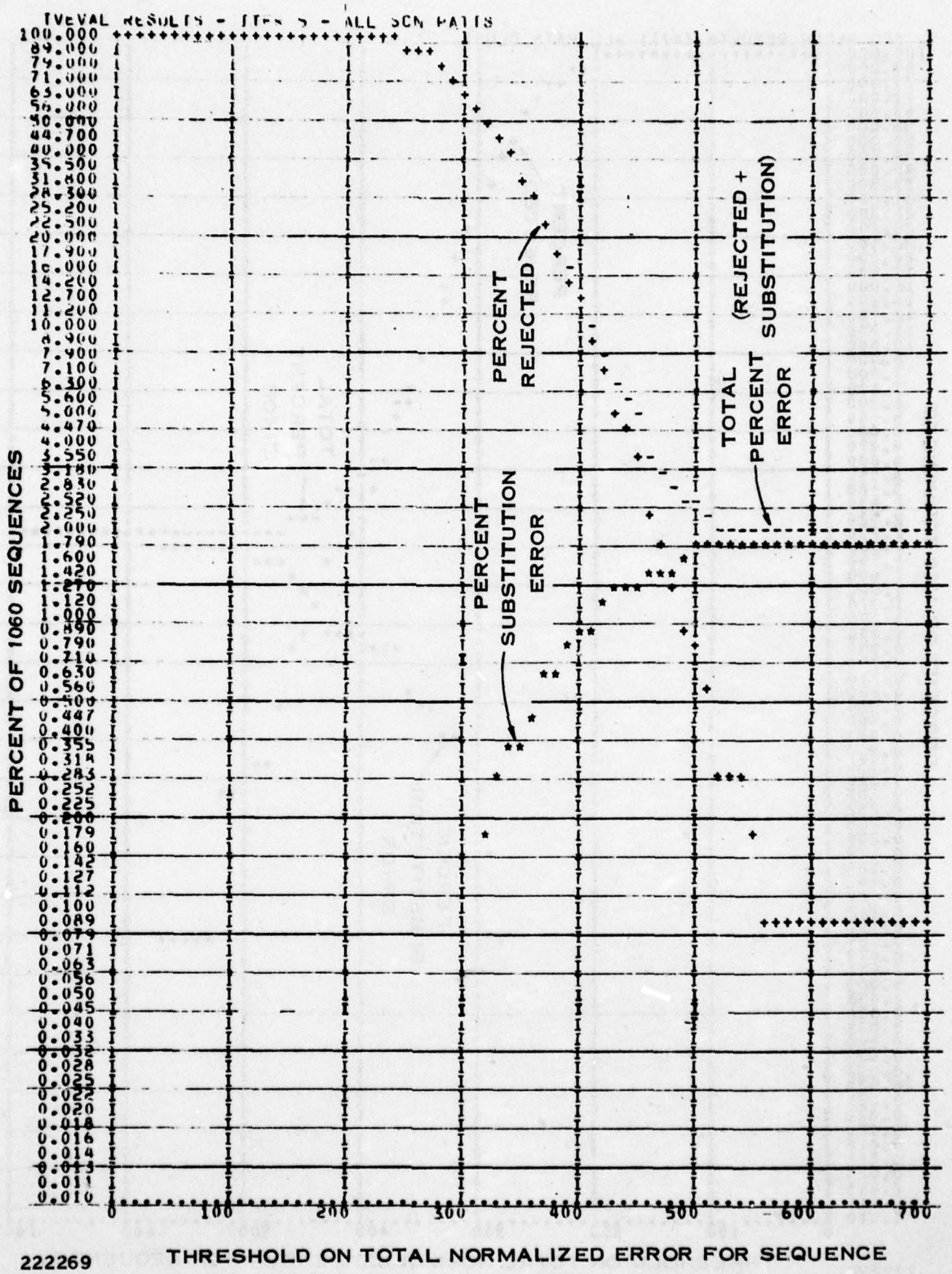


Figure 36. Run 5

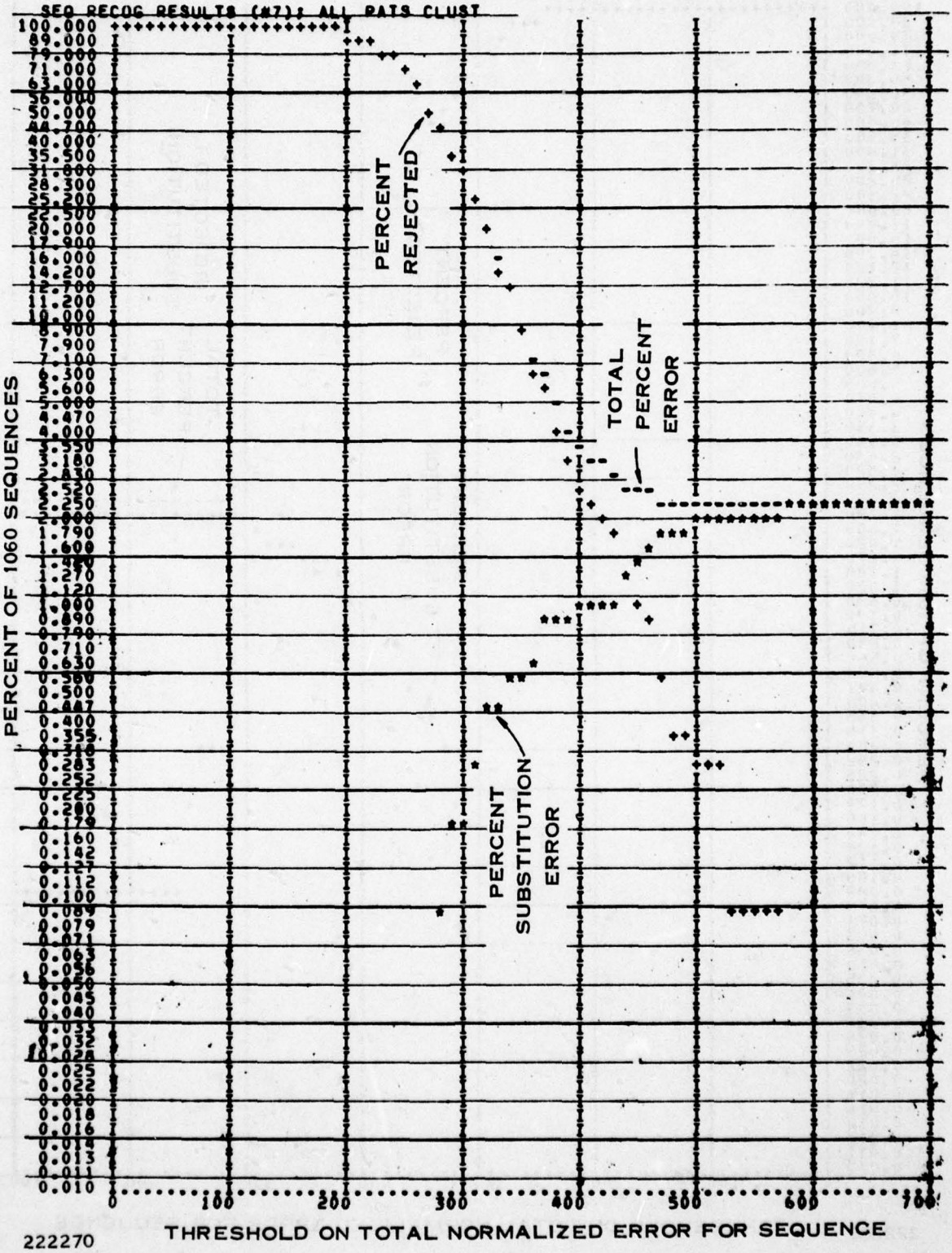


Figure 37. Run 7

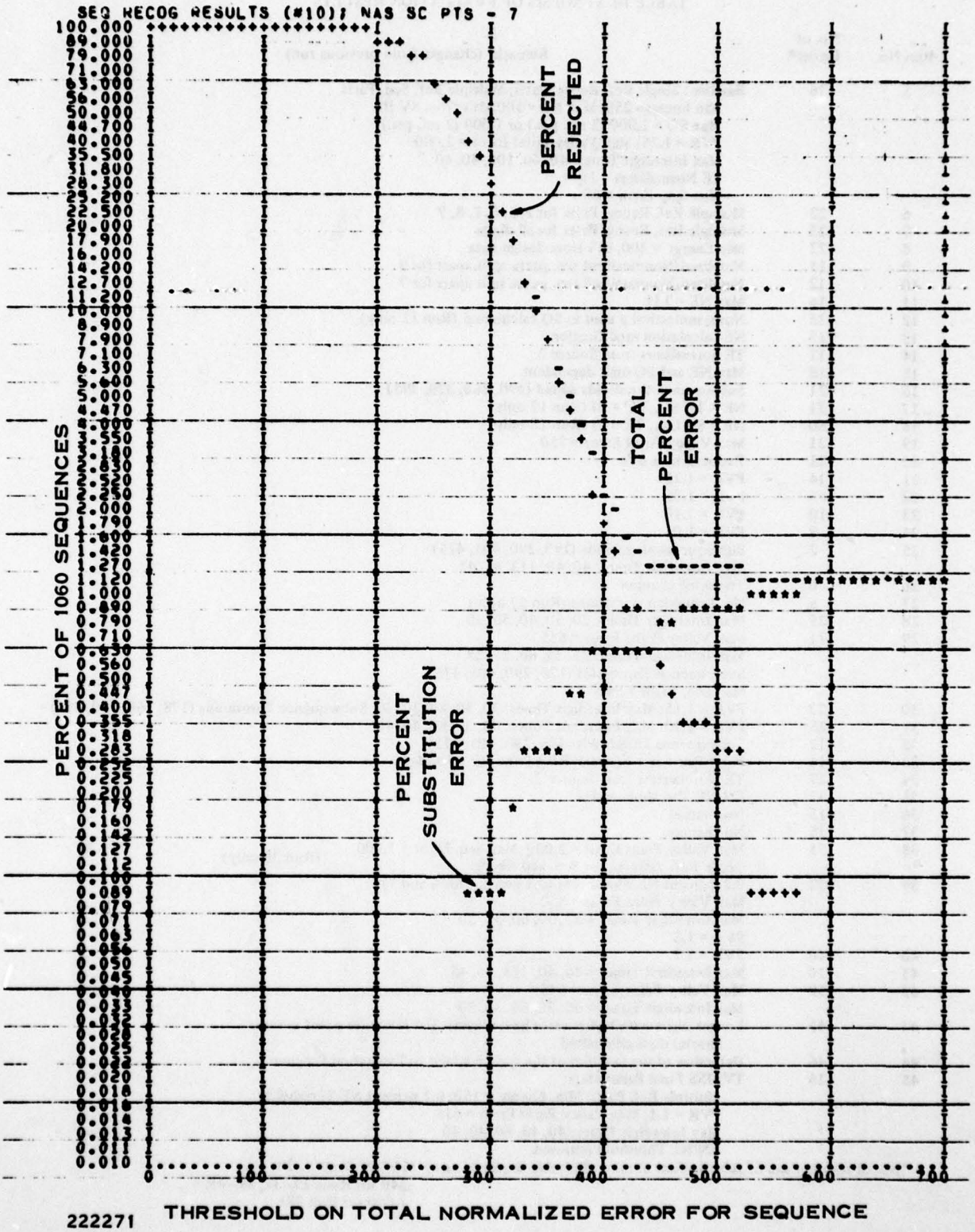


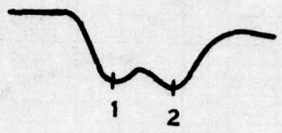
Figure 38. Run 10

TABLE 10. SYNOPSIS OF EVALUATION RESULTS

Run No.	No. of Errors*	Remarks (changes from previous run)
5	26	Baseline: Single Ref. Recog. Patts; Multiple Ref. Scn. Patts. Min Energy=250; Max NE = 500; dt's from SV III Max SQ = 2,000 (3 ref. pts.) or 1,000 (2 ref. pts.) PVR = 1.25; Max Valley Point Error = 2,000 Max Interdigit Times: 40, 40, 100, 40, 40 TE Normalizers = 1; Max Seq. Error = 600
6	22	Multiple Ref. Recog. Patts. for 2, 5, 6, 7, 8, 9
7	23	Multiple Ref. Recog. Patts. for all digits
8	22	Min Energy = 200; dt's from design data
9	15	Nasalized/Nonnasalized scn. patts. split apart for 9
10	12	Nasalized/Nonnasalized scn. patts. split apart for 7
11	16	Max NE = 131
12	15	Nonsymmetrical β used in SQ calculation (Run 12 only)
13	17	NE calculation modification
14	17	TE normalizers from Source 2
15	18	Max NE and SQ digit dependent
16	21	Subsequence thresholds added (190, 260, 320, 395)
17	21	NE = TE (i.e., SQ = 0) (Run 17 only)
18	60	NE = SQ (i.e., TE = 0) (Run 18 only)
19	21	Max Valley Point Error = 750
20	21	Reduced max dt's
21	14	PVR = 1.15
22	56	PVR = 1.35
23	10	PVR = 1.10
24	8	PVR = 1.05
25	2	Subsequence thresholds (190, 290, 405, 475) Max Interdigit Times: 40, 40, 113, 46, 45
26	0	Tweeking changes
27	6	No. forbidden transitions (Run 27 only)
28	25	Max Interdigit Times: 30, 30, 60, 30, 30
29	21	Max Valley Point Error = 615 Max Interdigit Times: 25, 25, 80, 25, 25 Subsequence Thresholds (178, 290, 405, 475) Max Seq. Error = 540
30	23	PVR = 1.15; Max Interdigit Times: 30, 30, 80, 30, 30; Subsequence Thresholds (178, 260, 320, 395)
31	15	PVR = 1.10; Max Interdigit Times: 30, 30, 80, 40, 40
32	12	Subsequence Thresholds (178, 290, 405, 475)
33	14	Point-Pair Error between RP #1 and RP #3 added
34	17	TE Normalizers from Source 3
35	17	EN/NE Threshold added
36	15	No changes
37	15	No changes
38	75	Max Valley Point Error = 2,000; Max Seq. Error = 1,000 Single Ref. Patterns for Scn. and Recog. (Run 38 only)
39	52	*0.2 percent NE Thresholds (0.1 percent for 6 and 7)* Max Valley Point Error = 850 Max Interdigit Times = 30, 30, 60, 30, 30 PVR = 1.2
40	40	PVR = 1.1
41	10	Max Interdigit Times = 40, 40, 113, 46, 45
42	39	Max Valley Point Error = 615; Max Interdigit Times = 30, 30, 60, 30, 30
43	46	Longer digits with higher error having same first reference point as same shorter digit eliminated
44	46	Detection of the location of the speech added to Evaluation Program
45	16	TVBISS Final Parameters: Multiple Ref. Patts; Min. Energy = 150; 0.1 percent NE Thresholds PVR = 1.1; Max Valley Point Error = 615 Max Interdigit Times: 40, 40, 80, 40, 40 EN/NE Threshold removed

*No. of rejects + no. of substitutions at max. allowable sequence error. (600 for Runs 5–28; 540 for Runs 29–37, 39–45; 1,000 for Run 38)

Using multiple reference scanning patterns necessitates using a lower PVR, since multiple dips in e_{RP_i} are more likely to occur in the scanning error functions, i.e.,



Reference point 2 would always be found, but point 1 will only be found with a small PVR.

Since valley point errors generally are not as low in speaker independent work, a smaller PVR is needed than for speaker dependent recognition.

Examples of PRVs used on various programs:

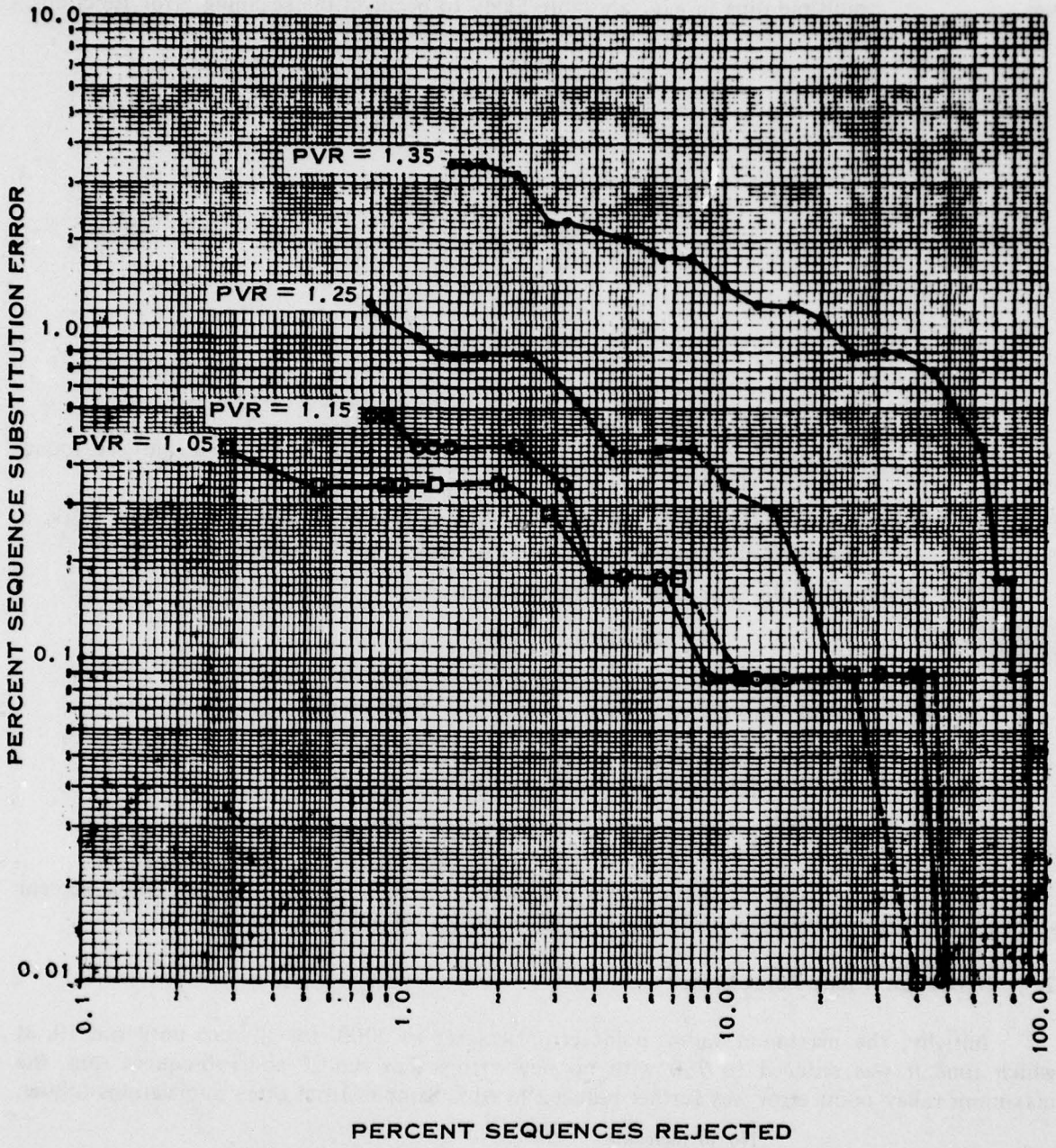
PVR	Application	
1.15	SVIII–Total voice speaker verification	Scans every other time sample
1.30	SVIII–Password detection	
1.25	Speaker independent word recognition	Internal programs
1.30	Speaker dependent word recognition	
1.50	CIC and BISS speaker verification (speaker dependent)	
1.10	TVBISS	

The results of runs 20 through 24 are summarized in Figure 39 which plots percent sequence substitution versus percent sequence rejection for several values of PVR.

2. Maximum Valley Point Error

Initially, the maximum valley point error was set to 2000, for all runs until run 19, at which time it was reduced to 750 with no new errors. For run 29 and subsequent runs, the maximum valley point error was further reduced to 615. Samples from other applications follow:

Maximum VPE	No. of Elements Per Vector	Application
200	51, 3-bit elements	SVIII–Password detection and digit recognition (speaker independent)
1,200	85, 3-bit elements and 68, 4-bit elements	Speaker independent and dependent word recognition
615	85, 3-bit elements and 68, 4-bit elements	TVBISS



222272

Figure 39. Substitution Versus Reject Rate for Several PVRs

3. dt Limits and Expected Values

Significant parameters in digit hypothesizing are the time restrictions for fitting reference points together in a sequence. The training data was intentionally chosen to illustrate short digits (digit position 1 with a following connected digit) and long digits (digit position 3, with no following connected digit). All digits except "seven" showed a definite bimodal distribution of distances between reference points. Therefore, for each of those nine digits, two sets of timing restrictions were supplied: one set defining short digits, and the other defining long digits. The timing restrictions are shown in Table 11.

TABLE 11. TIMING RESTRICTIONS FOR DIGIT HYPOTHESIS GENERATION

		Distance Between Ref. Pt. 1 and Ref. Pt. 2			Distance Between Ref. Pt. 2 and Ref. Pt. 3		
Digit		Expected	Minimum	Maximum	Expected	Minimum	Maximum
0	Short	16	2	28	12	2	14
	Long	19	2	30	21	15	38
1	Short	11	4	16			
	Long	21	17	40			
2	Short	6	2	16	17	4	23
	Long	8	2	18	29	24	50
3	Short	18	6	24			
	Long	32	25	56			
4	Short	16	6	20			
	Long	27	21	38			
5	Short	20	4	27			
	Long	32	28	50			
6	Short	8	2	10	7	2	20
	Long	14	11	30	10	2	20
7		12	4	24	11	4	28
8	Short	16	6	19			
	Long	23	20	40			
9	Short	20	8	23			
	Long	29	24	52			

Evaluation runs 1 through 7 used the dt values from the Speaker Verification III[7] study. Evaluation runs 8 and up used the dt values from Table 11. Changing dt had little effect, however, on performance.

The parameter for the "minimum expected dt" was set below all of the expected values of dt in Table 11 for all evaluation runs. Hence, $\hat{dt}^* \equiv \hat{dt}$.

4. **Time Deviation Weighting (β)**

No variation in the value of $\beta = 2$ was attempted. For run 12 only, the equation using β was modified to more lightly weight deviations for $dt < \hat{dt}$, but with an insignificant decrease in error rate and with a 40-percent increase in running time. It should be noted that the equation for the point pair error in TVBISS squares the deviation from the expected dt, rather than using just the magnitude of the deviation as was done for Speaker Verification III. This does not penalize small variations as much, but penalizes large variations more.

5. **Floor of the Valley Point Error (OFFSET)**

OFFSET = 100 for all evaluation runs.

6. **Minimum Average Energy Across Recognition Pattern (EN_{min})**

EN_{min} = 250 through run 7; EN_{min} = 200 for runs 8 to 44, and EN_{min} = 150 for the final run.

7. **Relative Weighting of Sequence Error (SQ) and Recognition Error (TE)**

The weighting constant, w_k, was set to 1 for all evaluation runs except 17 and 18. For evaluation run 17, the total normalized error (NE_k) was set equal to the normalized recognition error and for run 17, NE_k was set equal to the normalized sequence error. Error curves for run 17 were marginally worse than for $w_k = 1$. Error curves for run 18 were much worse than for $w_k = 1$. No other values were tried; hence, $w_k = 1$ in the final set of parameters.

8. **Recognition Error (TE) Normalization**

The normalizing constants are calculated from the expected values of TE_k as follows:

$$\text{normalizing constant}_k = \frac{10\ (TE_k / \text{no. columns in k})}{\sum_{i=0}^{9} (TE_i / \text{no. columns in i})}$$

Values for these normalizing constants are given in Table 12 using three different sources of data:

1. $E_e + E_s + E_a$ from Table 20 of the Speaker Verification III report[7]
2. Values of J_e for the number of reference patterns chosen for each digit (see Section IV, Clustering) from the design data
3. Values of TE for each of the digits in correctly recognized sequences in the test data set.

The values of the normalizing constants used during the evaluation runs are as follows:

Evaluation Run Numbers	TE Normalizers
1 to 12	All 1.0
13 to 33	Source 2 (JEs from clustering)
34 and 45	Source 3 (TEs from evaluation data set)

TABLE 12. TE NORMALIZING CONSTANTS

Digit	Source 1	Source 2	Source 3
0	1.039	1.022	0.998
1	1.170	1.089	1.225
2	0.708	0.837	0.878
3	1.085	1.016	1.021
4	0.853	0.719	0.681
5	1.182	1.292	1.195
6	0.705	0.870	0.788
7	1.052	0.945	0.977
8	1.007	1.133	1.008
9	1.200	1.076	1.229

Runs 11 and 13 were identical except for the TE normalizers. No significant difference was seen between runs. Similarly, runs 33 and 34 were identical except for the TE normalizers. Again, no significant difference was seen on six-digit sequence recognition, although the percent correct digit recognition (independent of syntactic constraints) improved from 89.4 to 91.7 percent.

9. Sequence Error (SQ) Thresholds

These thresholds are applied to the sum of the point-pair errors of sequences in order to limit the number of hypothesized digits. These values were determined directly from the test data and were set large enough so that all digits in correctly recognized digit sequences had sequence errors less than these thresholds, which follow.

Digit	0	1	2	3	4	5	6	7	8	9
SQ threshold	630	350	650	300	300	350	630	600	300	260

10. *Total Normalized Error Thresholds*

These thresholds are applied to hypothesized digits in order to limit the number of entries in the table of hypothesized digits. The thresholds on NE were determined from cumulative distribution plots such as that shown in Figure 40, which is a plot of the percentage of that digit in correctly recognized sequences [from the evaluation run (34) after TE normalizers were last changed] whose NE is less than the abscissa value. Note that this line is quite linear on the semi-log plot. Hence, although these lines are based on the test data, they would hold more generally. The values of NE at the 0.2- and 0.1-percent points for all digits are as follows:

Digit	0	1	2	3	4	5	6	7	8	9
0.2 percent	118	91	114	103	108	106	102	102	101	106
0.1 percent	128	97	123	110	116	113	110	109	107	114

Several evaluation runs used the 0.2-percent thresholds; however, higher normalized errors for some females required increasing these thresholds to the 0.1-percent values.

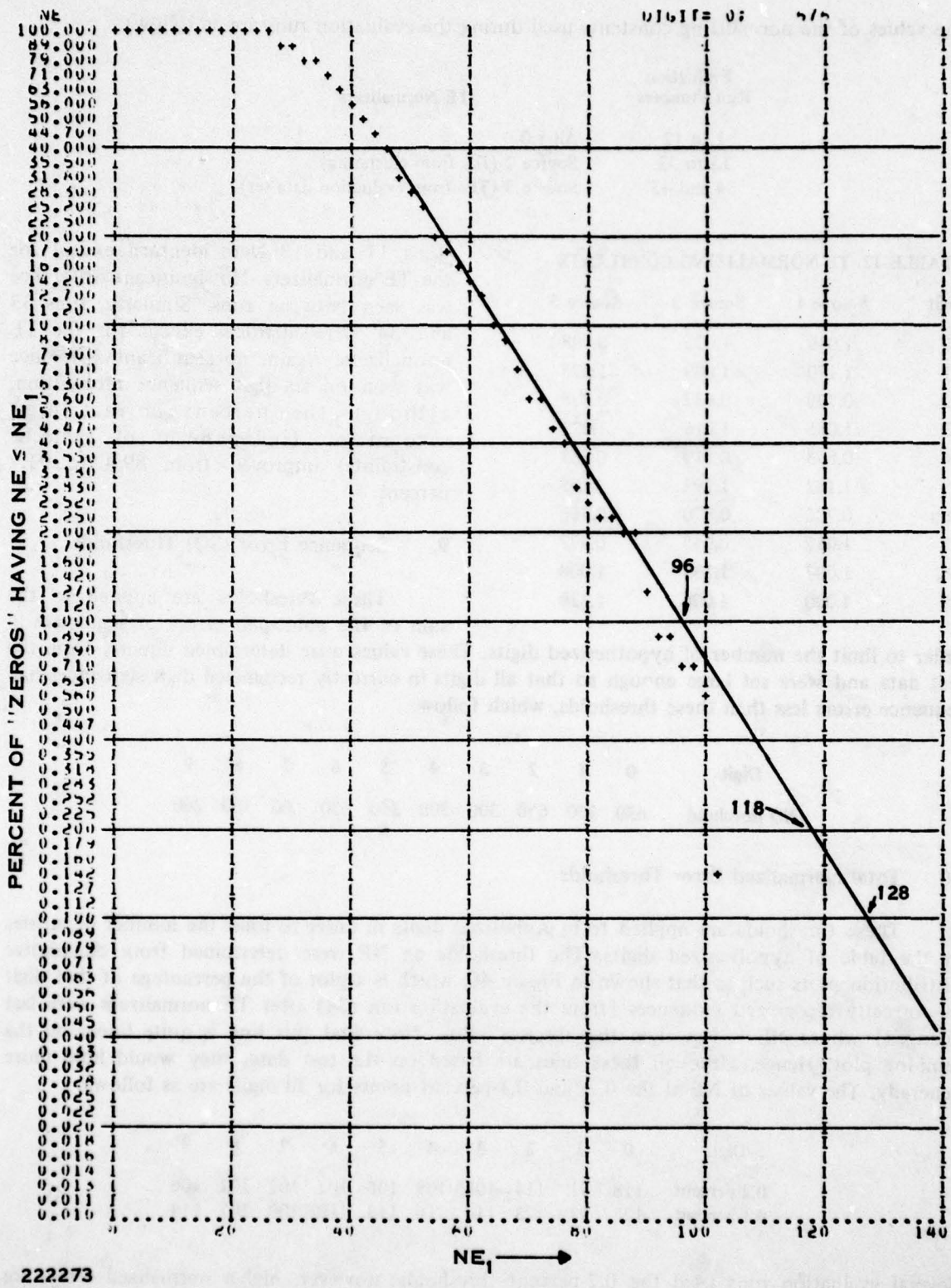


Figure 40. Determination for "Zero"

11. Syntactic Constraints

This topic is discussed in Section II.

12. Maximum Interdigit Times

Whereas, several previous thresholds were imposed in order to limit the number of entries in the table of hypothesized digits, both the maximum interdigit times and the maximum subsequence errors discussed next were imposed in order to limit the search time through the table. Since subjects are instructed to leave a short pause in the middle of the sequence, the following values are recommended:

Times (csec.)	**30**	**30**	**60**	**30**	**30**
Between digits	1,2	2,3	3,4	4,5	5,6

For demonstration purposes, times of 40, 40, 80, 40 and 40 are recommended to accommodate uninitiated users. During collection of the data used in the evaluation data base, values of 40, 40, 100, 40 and 40 were used, producing some errors during evaluation runs when shorter times were used.

13. Maximum Subsequence Errors

These error thresholds were selected from the test data using linear fits to the cumulative distribution curves as was done in selecting the NE thresholds. The subsequence thresholds used were at the 0.1-percent levels. Also listed below are the thresholds needed to accommodate the test data set. Note that since the searching is done from last digit to first, the subsequence thresholds apply from the sixth digit.

Digits	**5–6**	**4–6**	**3–6**	**2–6**
0.1-percent thresholds	190	260	320	395
Test data thresholds	178	290	405	475

14. Maximum Sequence Error Threshold

After the sequence with the minimum error is found, if one exists (which could actually be either the correct sequence or an incorrect one), the sequence error is compared to the maximum sequence error threshold. If it is below this threshold, it is accepted; otherwise, it is rejected. Figure 41 describes this decision. Notice the tradeoff between rejections and substitutions as the maximum sequence error threshold is shifted. This is also shown in Figure 42. Although the total error (substitutions + rejections) should be kept as low as possible, since this digit recognition is the front-end for a Speaker Verification system, it is preferable to have the bulk of the error be rejections, having the system prompt the user to repeat the sequence rather than to attempt a verification on an incorrect number as would be done on a substitution error.

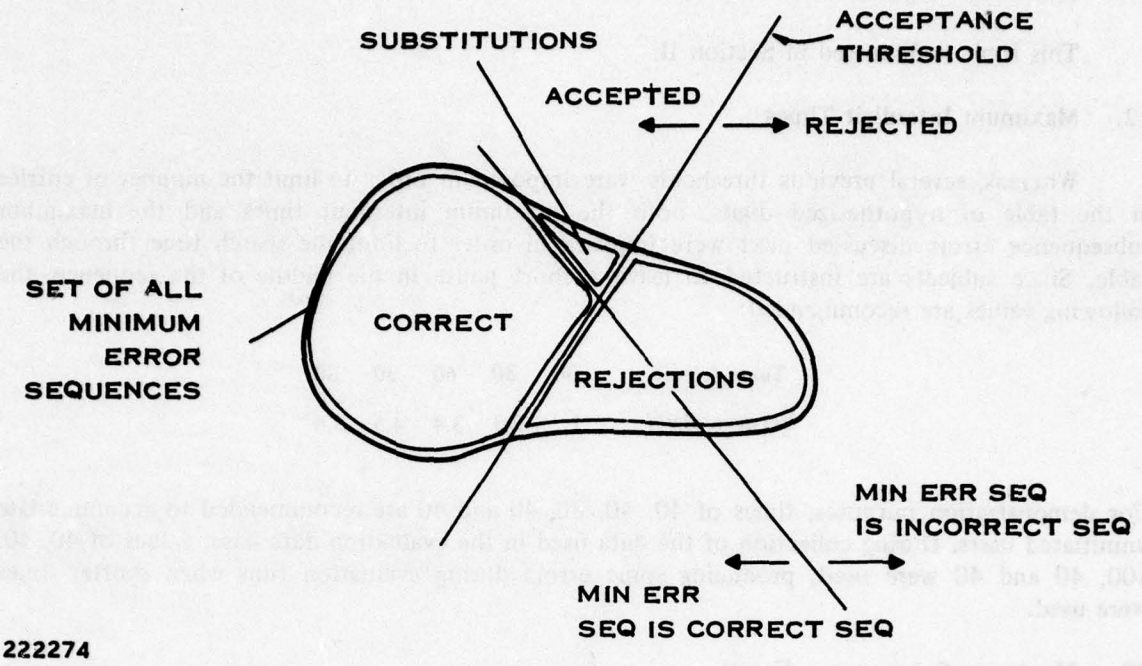


Figure 41. Sequence Error Threshold

A problem exists, however, in that females have higher NEs and, hence, higher total sequence errors, as seen by comparing reject rates in Figures 43 and 44 for sequence error thresholds between 200 and 350. At higher thresholds (~400), these curves become very much a function of the particular errors of a few problem speakers. For example, one male speaker had seven of the eight highest sequence errors, causing the shift right of the male curve around 400. Generally, a sequence error threshold set to yield low substitutions on Figure 42 will, on the average, yield a higher rejection rate for females. Specific females have even higher NEs and sequence errors that yield still higher rejection rates. Hence, instead of the threshold in the 400 to 420 range suggested by Figure 39, this threshold was set at 480.

B. SEQUENCE RECOGNITION TESTS

Two sets of recognition tests were done using all of the final parameters described in the previous section. The first test was using TVEVAL on the test data base of 1060 sequences from 106 speakers (100 of the 320 sequences were in this data base). Figures 42 through 44 were, in fact, results of this final evaluation run, which had 16 errors (1.5 percent): 11 rejections (1.0 percent) and 5 substitutions (0.5 percent).

Table 13 details the reasons for the correct sequences not being found for these 16 cases. Note that 9 of these 16 errors are due to too much time between digits, which is actually a speaker indoctrination problem rather than a program problem. The notations in parentheses on many of these sequences refer to the fact that the "end-of-speech" is assumed after the energy

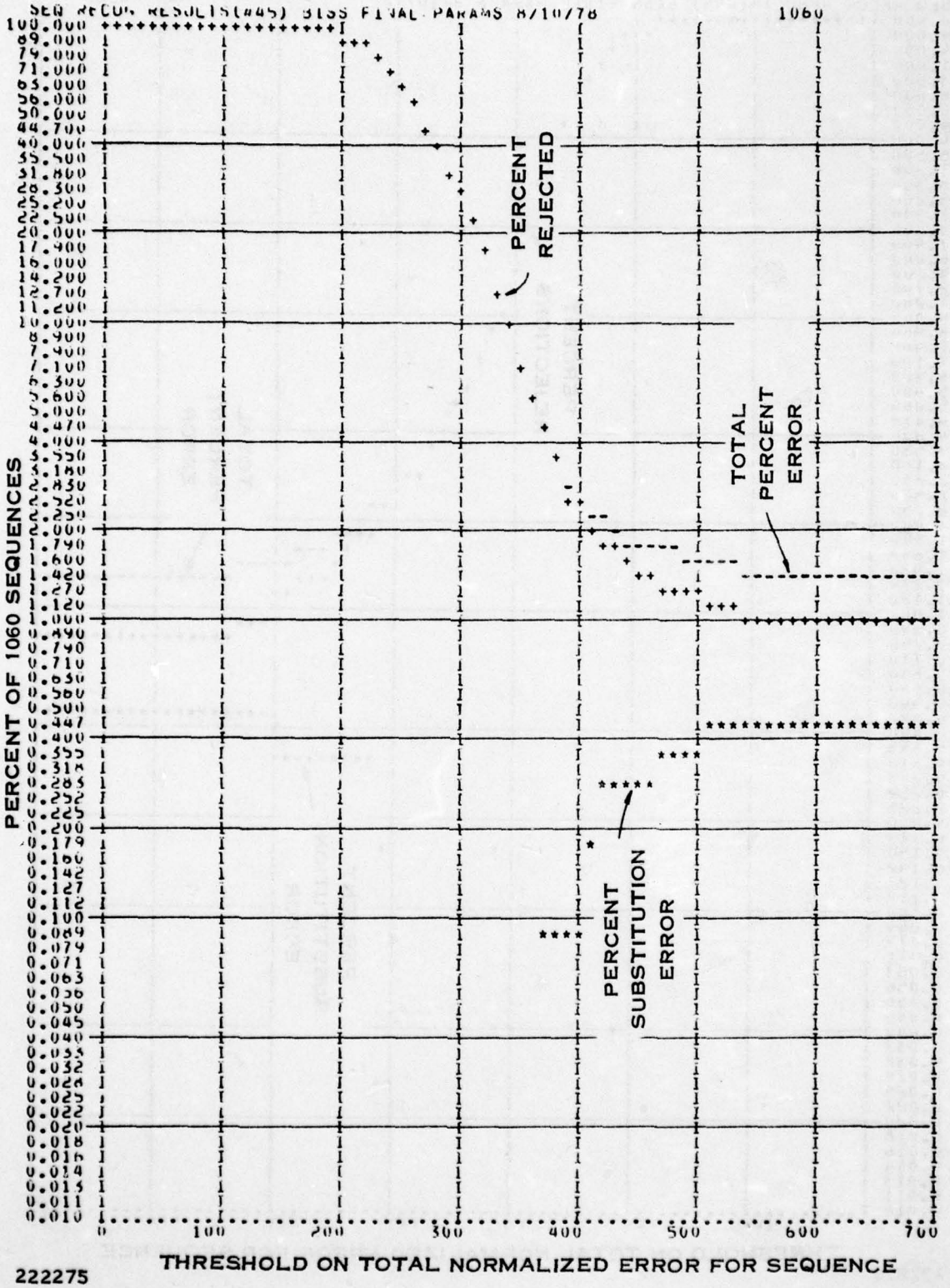
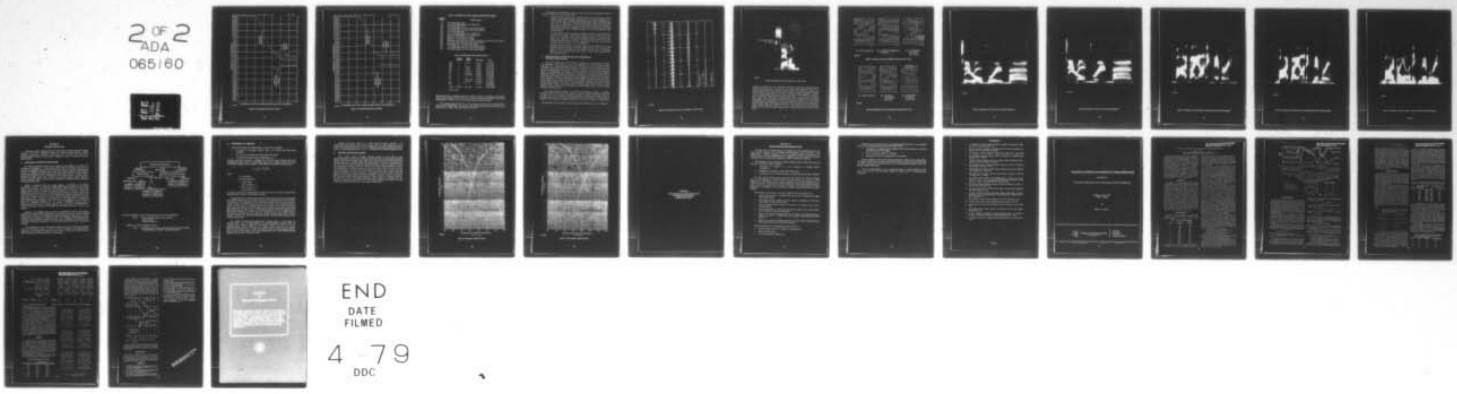


Figure 42. Final Evaluation Run Results–All Speakers

AD-A065 160 TEXAS INSTRUMENTS INC DALLAS F/G 17/2
TOTAL VOICE SPEAKER VERIFICATION.(U)
JAN 79 R L DAVIS, B M HYDRICK, G R DODDINGTON F30602-76-C-0329
UNCLASSIFIED RADC-TR-78-260 NL

2 OF 2
ADA
065160



END
DATE
FILMED
4 79
DDC

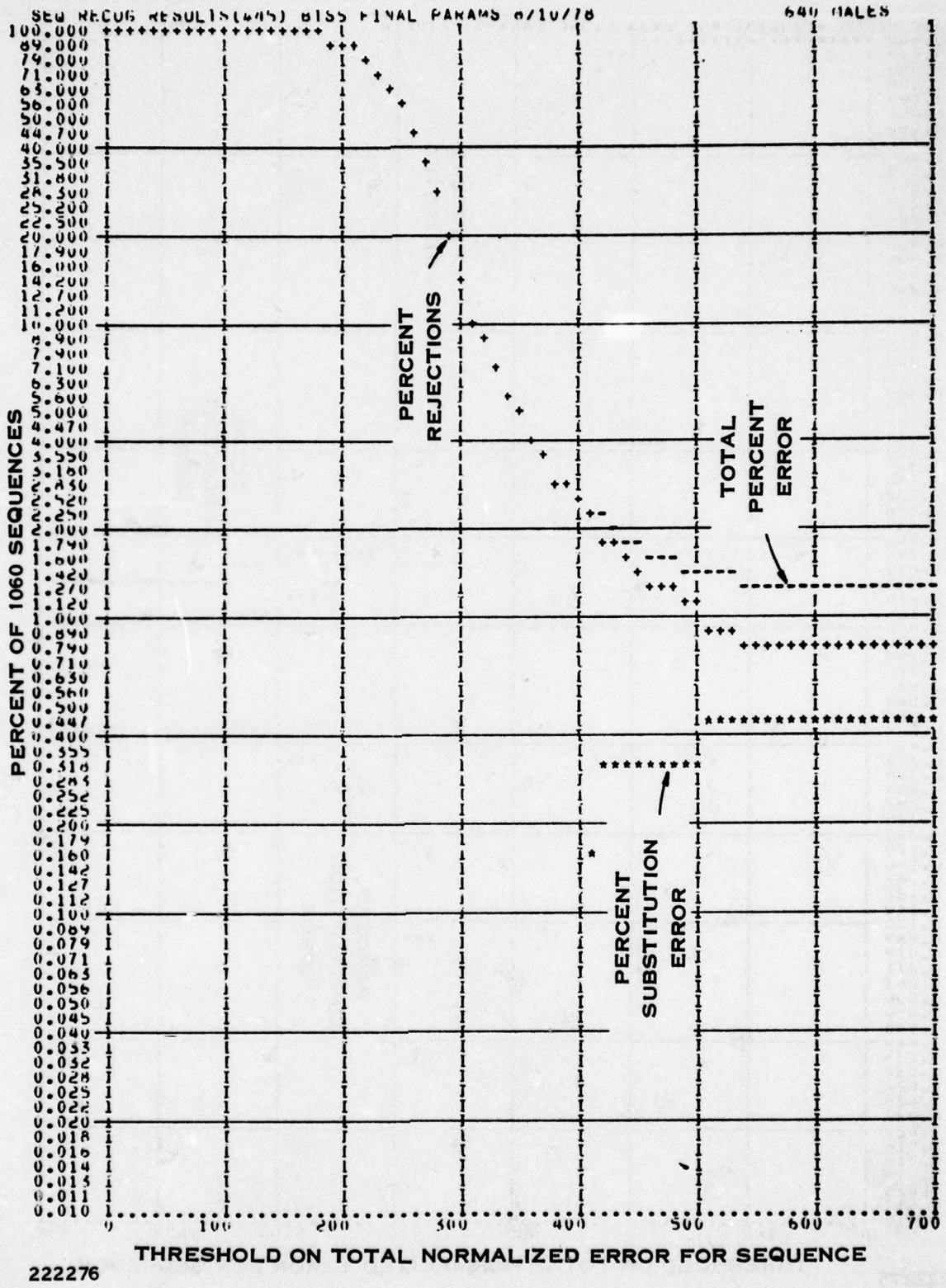


Figure 43. Final Evaluation Run Results—All Males

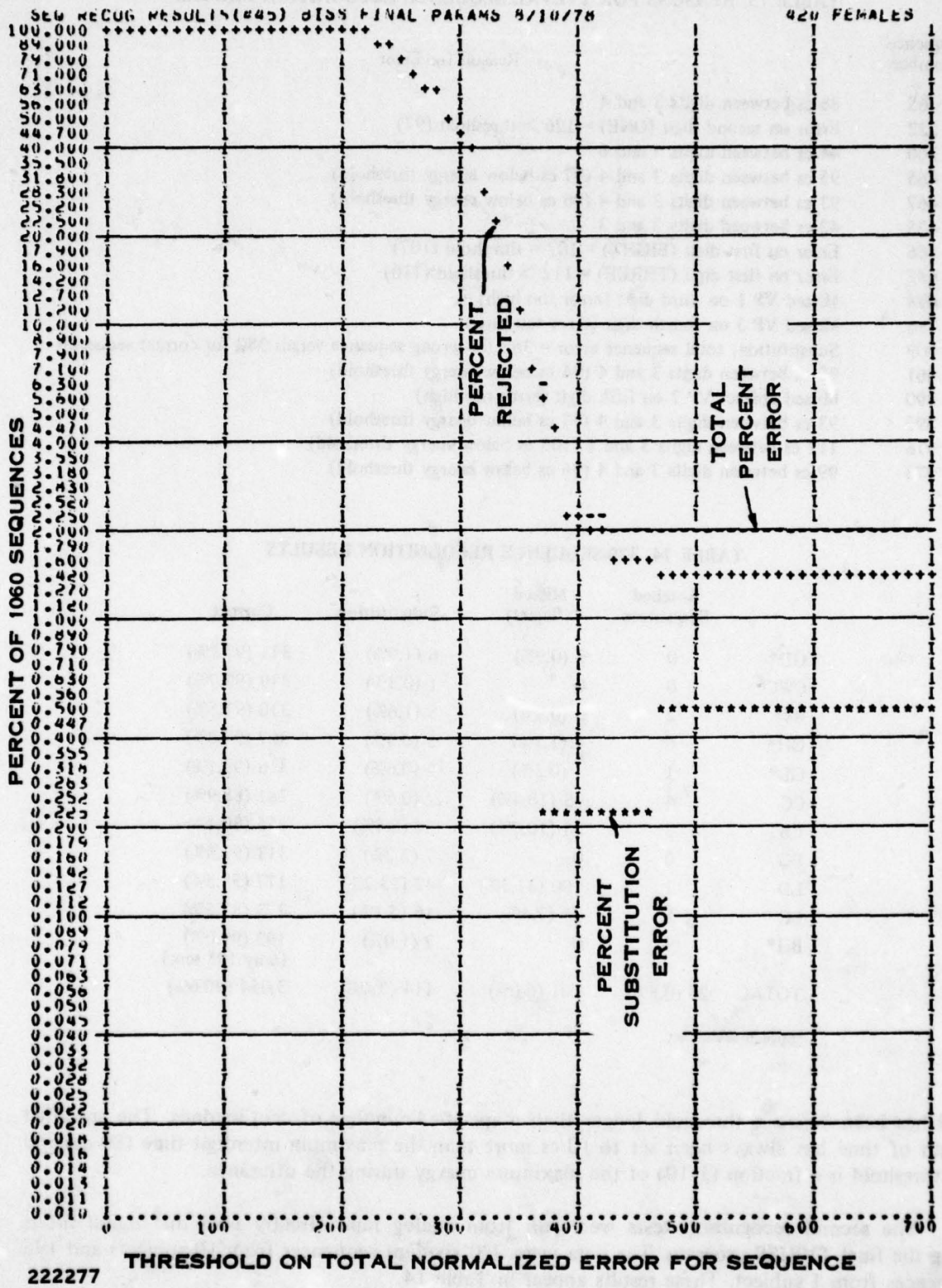


Figure 44. Final Evaluation Run Results—All Female

TABLE 13. REASONS FOR TVEVAL SEQUENCE RECOGNITION ERRORS

Sequence Number	Reason for Error
65	88 cs between digits 3 and 4
222	Error on second digit (ONE) = 126 > threshold (97)
260	44 cs between digits 5 and 6
465	95 cs between digits 3 and 4 (97 cs below energy threshold)
467	93 cs between digits 3 and 4 (96 cs below energy threshold)
425	42 cs between digits 2 and 3
566	Error on first digit (EIGHT) = 107 = threshold (107)
687	Error on first digit (THREE) = 112 > threshold (110)
698	Missed VP 1 on third digit (error too high)
796	Missed VP 3 on fourth digit (error too high)
809	Substitution: total sequence error = 368 for wrong sequence versus 380 for correct sequence
861	95 cs between digits 3 and 4 (94 cs below energy threshold)
890	Missed correct VP 2 on fifth digit (error too high)
895	93 cs between digits 3 and 4 (97 cs below energy threshold)
976	112 cs between digits 3 and 4 (105 cs below energy threshold)
978	99 cs between digits 3 and 4 (96 cs below energy threshold)

TABLE 14. 320-SEQUENCE RECOGNITION RESULTS

	Botched Sequences	Missed (Reject)	Substitution	Correct
GD*	0	3 (0.9%)	6 (1.9%)	311 (97.2%)
CWC*	0	0	1 (0.3%)	319 (99.7%)
BS*	2	3 (0.9%)	5 (1.6%)	310 (97.5%)
GH*	6	4 (1.3%)	3 (0.9%)	307 (97.8%)
GL*	1	1 (0.3%)	2 (0.6%)	316 (99.1%)
CC	4	33 (10.4%)	2 (0.6%)	281 (88.9%)
FB	3	34 (10.7%)	28 (8.8%)	255 (80.4%)
FG	0	0	7 (2.2%)	313 (97.8%)
LD	1	100 (31.3%)	42 (13.2%)	177 (55.5%)
LC	8	23 (7.4%)	16 (5.1%)	273 (87.5%)
BH*	1	0	2 (1.0%)	192 (99.0%) (only 195 seq.)
TOTAL	26 (0.8%)	201 (6.0%)	114 (3.4%)	3,054 (90.6%)

*Speech researcher.

level has been below a threshold longer than a specified number of centiseconds. The specified length of time has always been set to 10 cs more than the maximum interdigit time (80 cs) and the threshold is a fraction (1/10) of the maximum energy during the utterance.

The second recognition tests were run from analog tape directly into the digital filters using the final TVBISS program. The data were 320 six-digit sequences from 10 subjects and 195 sequences from 1 subject. These results appear in Table 14.

The reasons for the disparities in these test results need further investigation. However, some of the suspected reasons are listed below.

1. The data from the speakers in the TVEVAL data set were extracted from near the middle of data from the speaker verification data base that was collected over a 3-month period. Hence, although these speakers were not speech researchers, they were used to speaking into a microphone. In addition, these data were collected on-line with an older version of the total-voice program, allowing the speakers to have adapted somewhat to knowing how to say the sequences to have them recognized. Note that on the 320-sequence data set, the speech researchers had quite good performance.

2. The environment in which the data were collected was different. The TVEVAL data set was collected inside a speech booth, whereas the 320-sequence data set was collected in a computer room environment with background line-printers, card readers, fans, talking, etc. In addition, the TVEVAL data was edited so that only the 6-digit utterance and the 20 to 30 cs preceding and following the utterance was included, which was not the case with the other data.

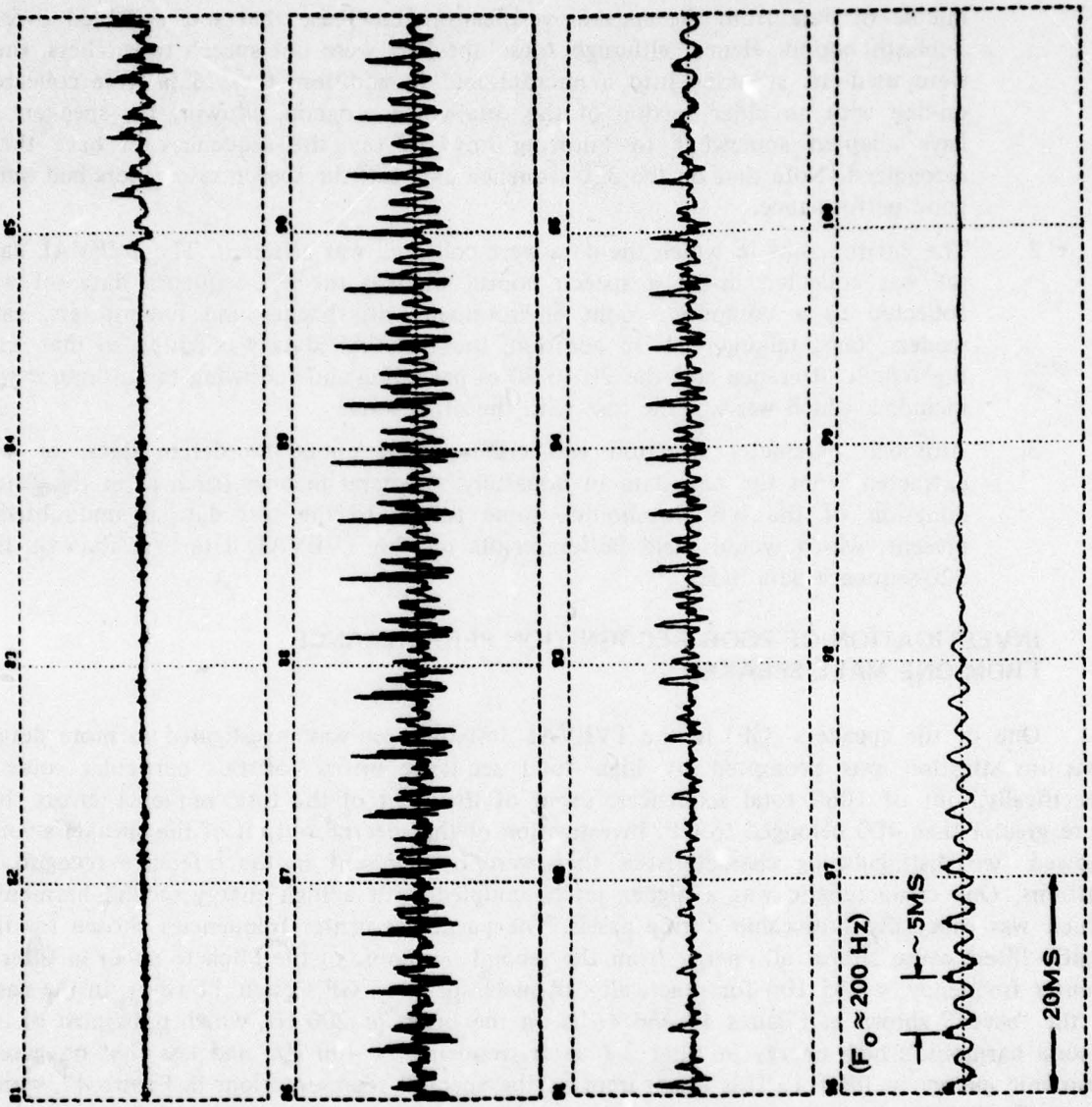
3. Although parameter selection was either based upon the design data, or was extracted from the test data in hopefully a general manner (such as in the determination of the NE thresholds), some tuning to the test data is undoubtedly present, which would yield better results on the TVEVAL data base than on the 320-sequence data base.

C. INVESTIGATION OF POOR RECOGNITION PERFORMANCE FROM ONE MALE SPEAKER

One of the speakers (GF) in the TVEVAL test data set was investigated in more detail. This investigation was prompted by high total sequence errors for this particular subject. Specifically, out of 1060 total sequences, seven of the eight of the total sequence errors that were greater than 400 belonged to GF. Investigation of the spectral output of the speaker's voice showed two distinguishing characteristics that were not present in the reference recognition patterns. One characteristic was a higher pitch, coupled with a high energy second harmonic, which was especially noticeable during nasals. The particular center frequencies chosen for the digital filters cause almost all energy from the second harmonic of the pitch to occur in filter 1 (center frequency $\approx$ 285 Hz) for practically all male speakers. GF's pitch, however, in the nasal in the "seven" shown in Figures 45 and 46 is on the order of 200 Hz, which puts most of the second harmonic's high energy in filter 2 (center frequency $\approx$ 400 Hz) and less first or second harmonic energy in filter 1. This is apparent in the spectral representations in Figure 47, which shows the input recognition pattern, the closest reference recognition pattern, and the absolute difference recognition pattern.

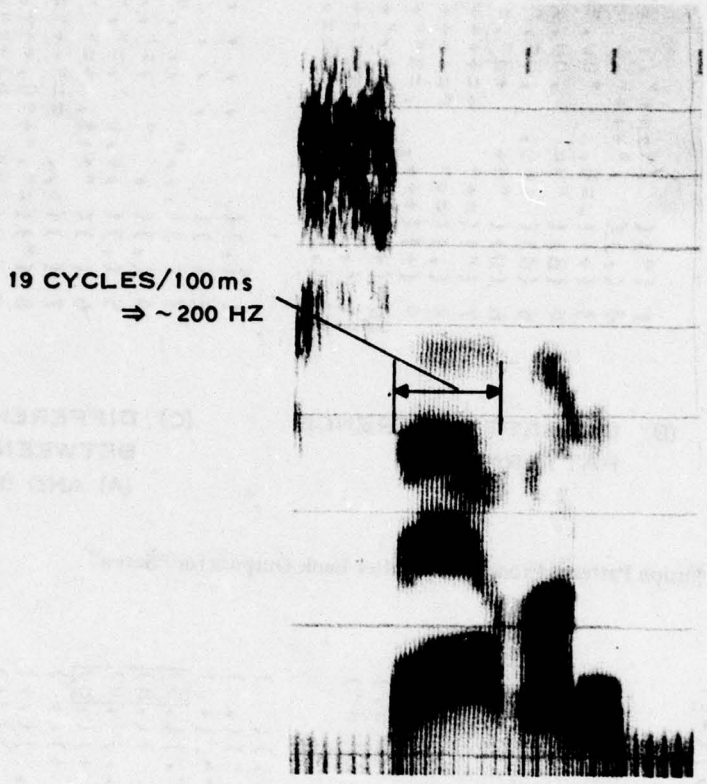
The second difference appears in the value of the regression coefficient c_2 especially during the vowels. An example of this is shown in Figure 48, showing recognition patterns for the digit "five." The c_2 regression coefficient indicates how rapidly speech energy falls with increasing frequency (how much spectral rolloff). During voiced sounds, the spectral slope is affected by the shape of the glottal driving pulse.* The normal 6-dB/octave rolloff of the

*The relationship of glottal pulse shape to frequency response is discussed fully in Section 6.241 of Flanagan.[16]



222278

Figure 45. Time Waveform for "Seven" From Sequence "152374" for GF



222279

Figure 46. Spectrogram for "Seven" From Sequence "152374" for GF

frequency spectrum is due to the combination 12-dB/octave rolloff due to the glottal pulse and 6-dB/octave preemphasis due to radiation. Thus, a 6-dB/octave preemphasis on the input to a spectrum analyzer should serve to merely flatten the spectrum. Figure 49 shows a sample of the subject's speech with no preemphasis, and Figure 50 shows a sample with preemphasis. It appears that the subject's glottal pulse has less than a 12-dB/octave rolloff, resulting in a larger c_2 regression coefficient. While listening to the analog recordings, it was noticed that this speaker used a higher than normal effort in speaking. Subsequent spectrograms made using another (female) subject with three different levels of effort are shown in Figures 51, 52, and 53. These spectrograms show the same effect on the spectral shaping with increased effort as that shown with the original subject. This increased spectral level of the formants with increased intensity or effort is explained more fully in Section 2.3 of Fant.[17] The conclusion, then, is that this second problem can be avoided by proper indoctrination of subjects with respect to proper speech effort.

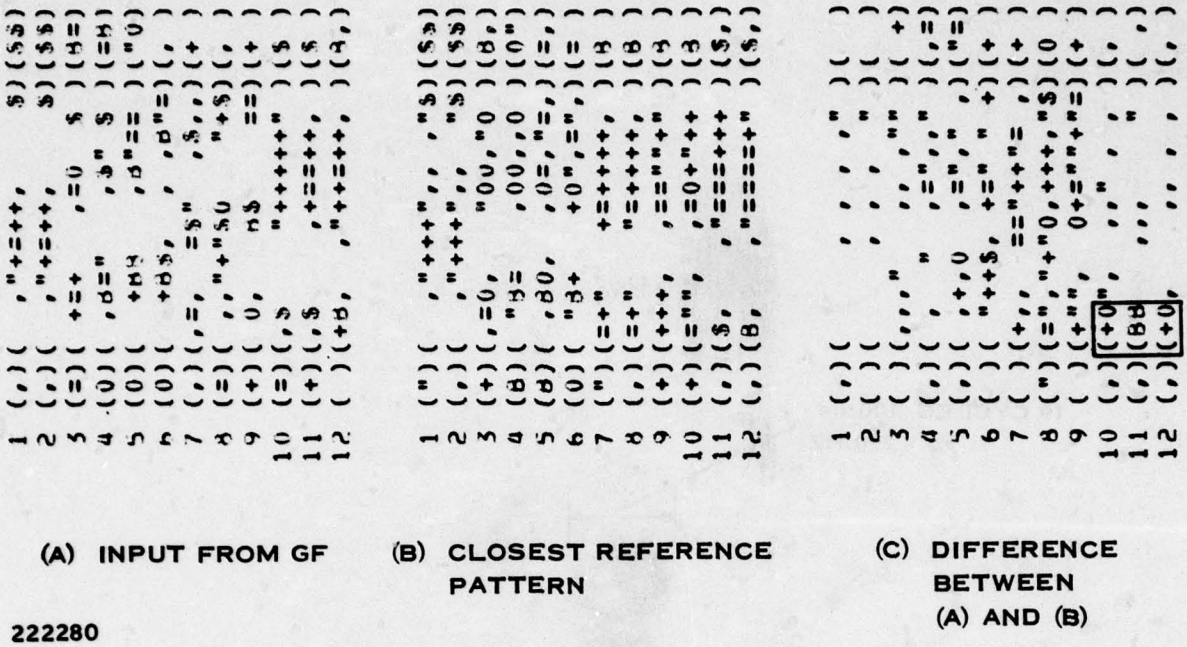


Figure 47. Recognition Patterns From Digital Filter Bank Outputs for "Seven"

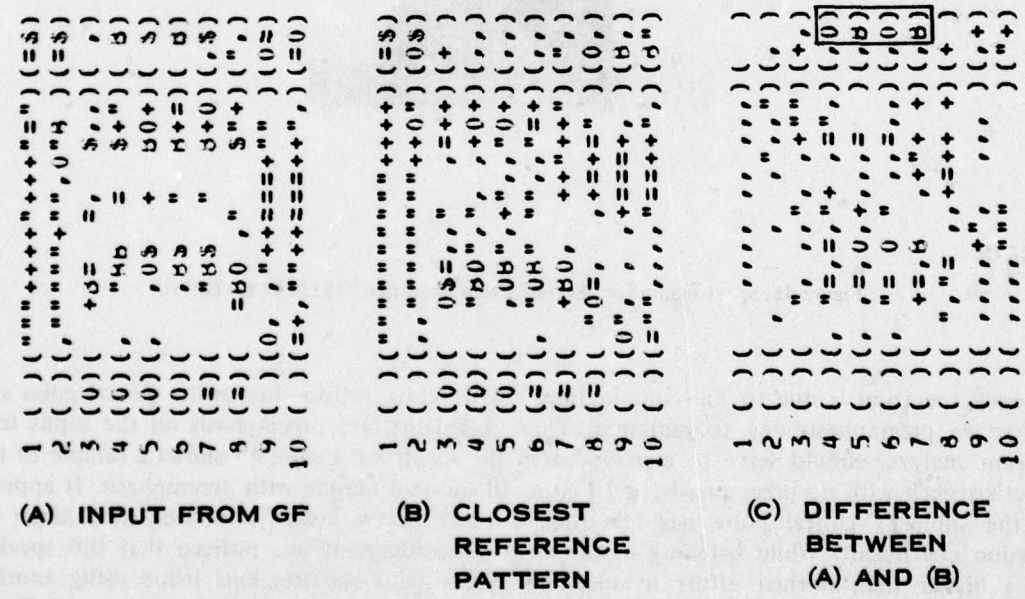
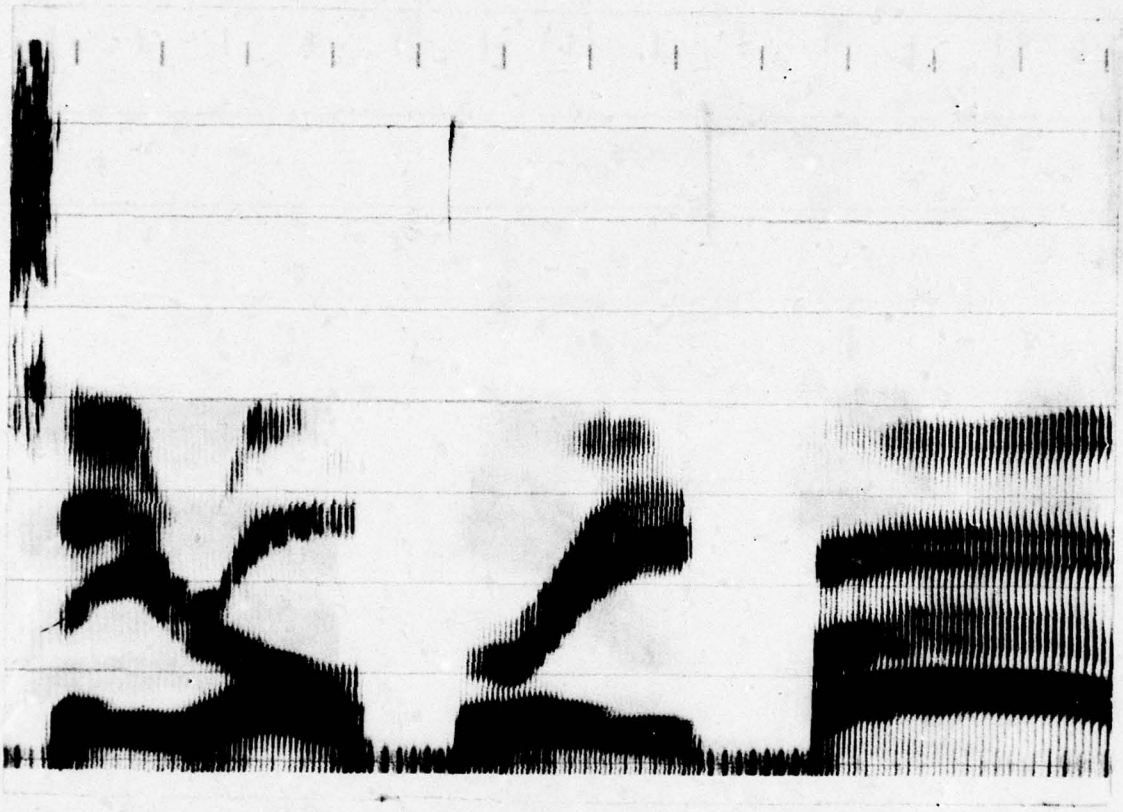


Figure 48. Recognition Patterns From Digital Filter Bank Outputs for "Five"



222282

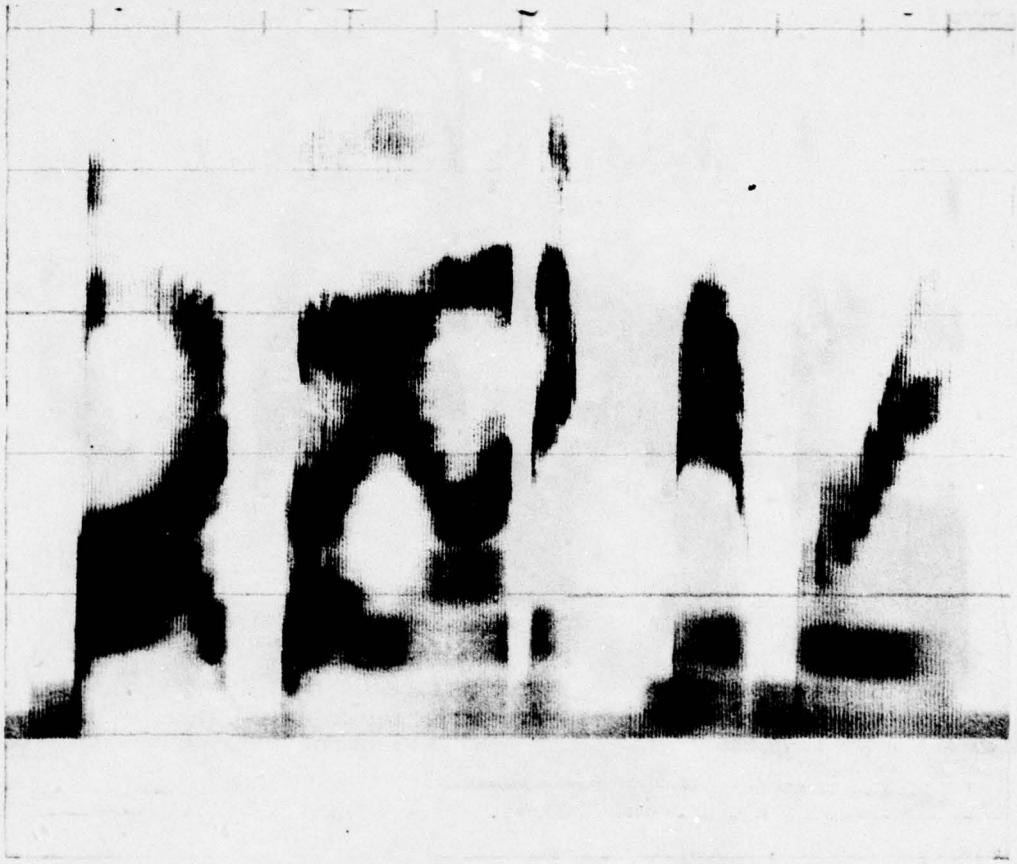
Figure 49. Spectrogram of "035" From GF's Data (Without Preemphasis)

222283

Figure 50. Spectrogram of "035" From GF's Data (With Preemphasis)

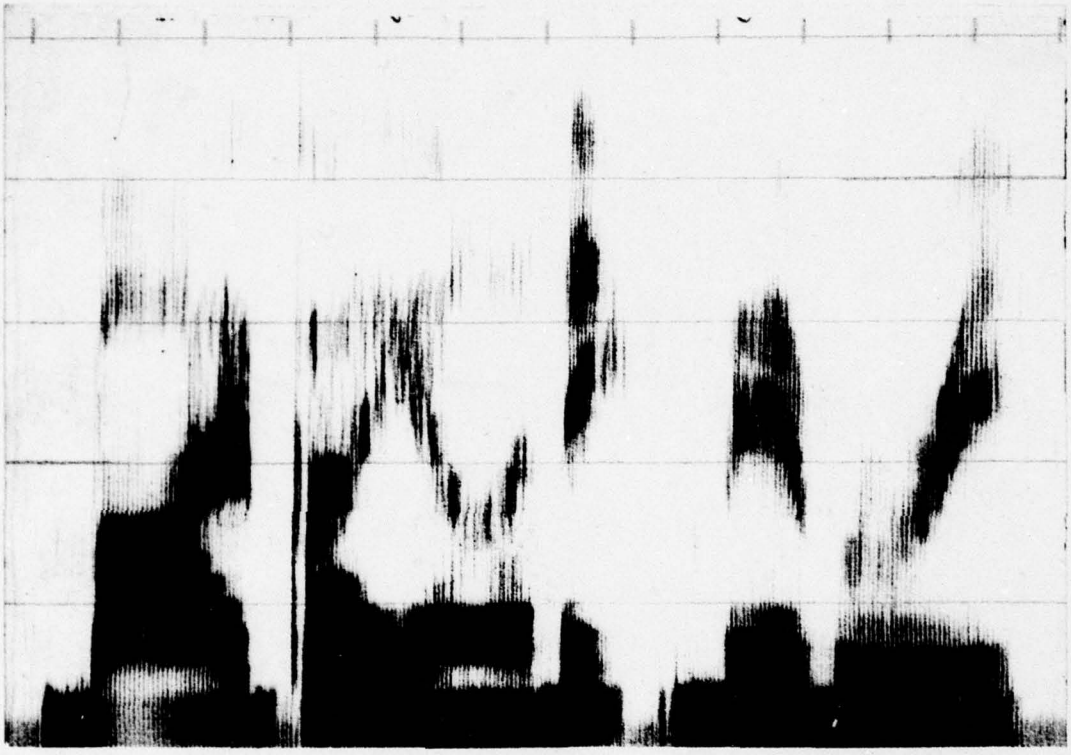
222284

Figure 51. Spectrogram of "My Bionic Memory" With High Speech Effort (With Preemphasis)



222285

Figure 52. Spectrogram of "My Bionic Memory" With Moderate Speech Effort (With Preemphasis)



222286

Figure 53. Spectrogram of "My Bionic Memory" With Low Speech Effort (With Preemphasis)

SECTION VI
SPEAKER VERIFICATION

Although the primary emphasis during this contract was on the digit sequence recognition front-end to the speaker verification program, the delivered program included a speaker verification component as described in this section. The same input data used in the speaker-independent fashion for the sequence recognition is used for verifying the speaker's claimed identity.

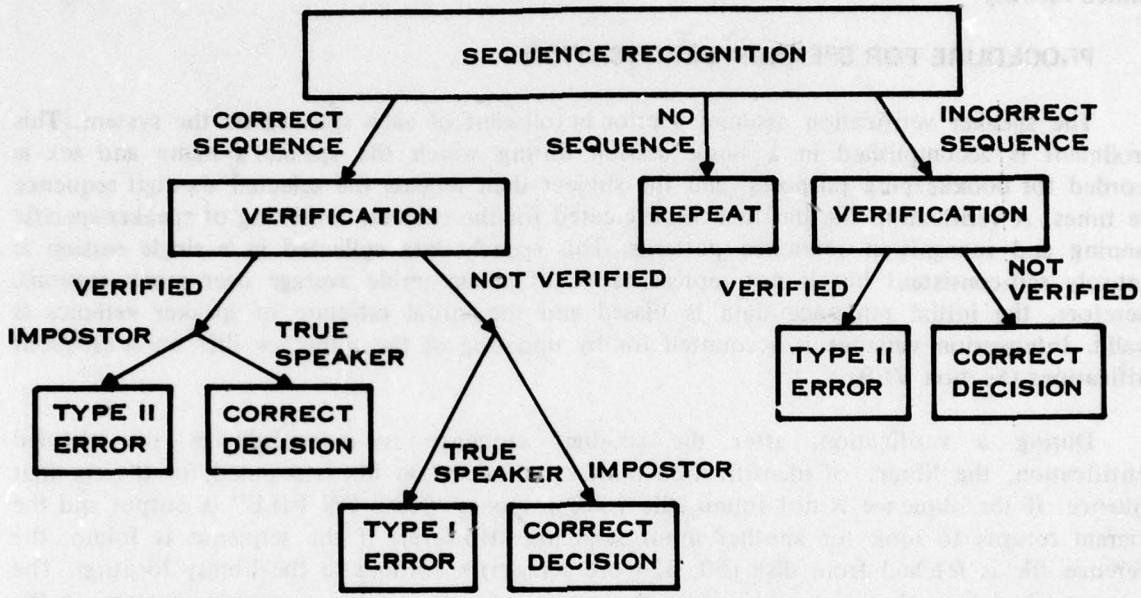
A. PROCEDURE FOR SPEAKER VERIFICATION

The speaker verification assumes a prior enrollment of each speaker on the system. This enrollment is accomplished in a single session during which the speaker's name and sex is recorded for bookkeeping purposes, and the subject then repeats the selected six-digit sequence five times. A verification file has then been created for the subject, consisting of speaker-specific scanning and recognition reference patterns. This speech data collected in a single session is relatively self-consistent but is not representative of an ensemble average over many sessions. Therefore, the initial reference data is biased and the initial estimate of speaker variance is invalid. Intersession variance is accounted for by updating of the reference files on subsequent verifications (Section VI.B).

During a verification, after the six-digit sequence is recognized as the claimed identification, the library of identification numbers that are on file is scanned for the six-digit sequence. If the sequence is not found, the voice response "NOT ON FILE" is output and the program returns to look for another input sequence. However, if the sequence is found, the reference file is fetched from disk (60, 32-word sectors) according to the library location. The input speech data is then rescanned using the speaker-specific reference scanning patterns in the region of ±10, 10-msec time samples around the speaker independent reference point locations found during sequence recognition. The recognition error is recomputed for all six digits using speaker-specific reference recognition patterns. If the sum of the recognition errors is less than or equal to a "verification threshold," the voice response "VERIFIED" is output, and the speaker-specific reference files are updated. Otherwise, the voice response "NOT VERIFIED" is output. In both cases, all arrays are reinitialized and the program returns to look for a new sequence.

Presently, all verification trials are treated as independent; however, both sequence recognition and speaker verification performance can obviously be aided by a sequential decision strategy since the trials are not independent. Sequential hypothesis testing such as the Wald Sequential Test is described, for example, in section 3.5 of Fukunaga.[18] Specific application of sequential hypothesis testing to speaker verification is given by Doddington and Hydrick.[4] Such a procedure is considered mandatory for any operational authentication system using personal attributes.

The probabilities of Type I (true speaker rejection) and Type II (impostor acceptance) errors now must contain terms to account for sequence recognition errors (although this is analogous to misread badges or keyboards in prior systems for inputting the claimed identity). This is illustrated in the simple figure on the following page.



The Type I probability of error is then (ignoring the forms of the distribution),

$$P_e(I) = P(\text{not-verified}|\text{true speaker, correct sequence}) \cdot P(\text{true speaker}) \cdot P(\text{correct sequence})$$

Similarly, the Type II probability of error is

$$P_e(II) = P(\text{verified}|\text{impostor, correct sequence})\, P(\text{impostor})\, P(\text{correct sequence} + P(\text{verified}|\text{incorrect sequence}) \cdot P(\text{incorrect sequence})$$

B. REFERENCE FILE UPDATING

There are three reasons for speaker-specific reference files to be updated.

1. To account for intersession variance, which is not present with single session enrollment
2. To account for the adaptation of the speaker to the system
3. To account for long-term speaker changes (aging, etc.).

The last of the above reasons has a negligible effect. The first reason, however, needs to be accommodated quickly, with the second reason needing a more moderate updating rate. Thus, the following updating formula is applied to both scanning and recognition reference patterns.

$$R' = \frac{\alpha}{16} X + \frac{(16 - \alpha)}{16} R$$

where

R = old reference

R′ = updated reference

X = current input

$$\alpha = \begin{cases} 4 \text{ for updates } 1-4 \\ 2 \text{ for updates } 5-12 \\ 1 \text{ for updates } >12 \end{cases}$$

If a person is verified, his expected error and recognition patterns are updated. Scanning pattern updating, however, requires a more judicious procedure.

Scanning patterns are five time-sample windows in the spectral data centered around points of maximum spectral change. Often, though, one side of the scanning pattern is more consistent. One example is the /zI/ transition which is sometimes pronounced as /zi/ by the same person. Another example is the final reference point in zero, two, or three, which would always have the vowel in the first part of the reference pattern, but if followed by a six or seven, sometimes would contain the sibilant form in the second half of the pattern and sometimes would contain silence, depending upon the existence of a short silence segment between the digits. Since the filter banks are sampled on a 10 msec basis, there is always up to a ±5 msec variation from the ideal reference point time. This time variation in a reference scanning pattern having a "stability imbalance" would tend to drift in the more stable direction, eventually drifting completely away from the reference point.

One method of anchoring the speaker-specific scanning points is to only update the scanning patterns when the speaker-specific and speaker-independent reference points are coincident. This, however, requires that the speaker's patterns be close to one of the speaker-independent patterns. Otherwise, his reference pattern would never be updated. Therefore, the procedure used is to update the scanning pattern if the speaker-specific reference point is within ±2 time samples of the speaker-independent pattern, which although allowing updating, would still limit the amount of drift.

Another alternative might be to update using the pattern extracted at the speaker-independent reference point if the two points are ±1 or 0 time samples apart, and to update with the pattern extracted at +1 (or −1) sample from the speaker-independent reference point if the speaker-dependent point is +2 (or −2) samples from the speaker-independent point.

C. SPEAKER VERIFICATION TESTING

Only a limited amount of speaker verification testing was done using the large speaker-verification data base that was collected. Ninety-eight of the 106 speakers were enrolled. Ninety-four (53 male, 41 female) Type I trials were run and 946 (542 male, 404 female) casual impostor trials were run from 96 speakers. Each speaker said each of the ten 6-digit sequences in his text at least once. If the sequence was correctly recognized, a verification attempt was made. The results appear in Figures 54 and 55. The number of Type II trials was sufficient to yield a smooth impostor curve; however, this was not true of the Type I trials. The equal error rates were approximately 4 percent for males and 5 percent for females. This is clearly not enough testing to determine the actual error rates for the speaker verification. However, these two figures show the need to make the verification thresholds a function of the sex of the speaker (which is known). Earlier speaker-verification work (Figures 30 and 31 of Speaker Verification III, reference 7) shows that additional performance can be gained by normalization by the speaker's expected error, which the Total Voice program calculates and updates with each verification, storing the updated value in the speaker's reference file.

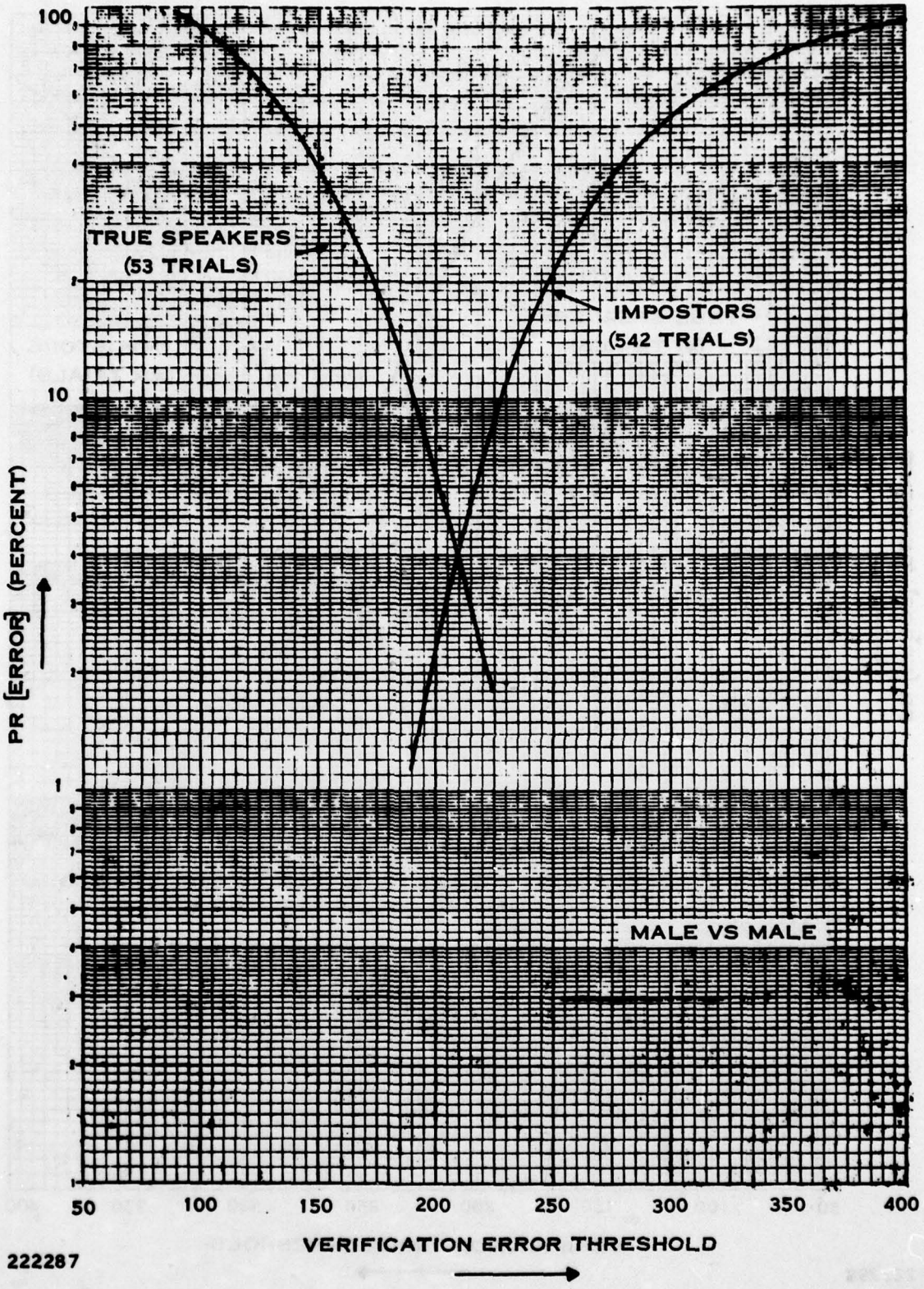


Figure 54. Male Speaker Verification Results

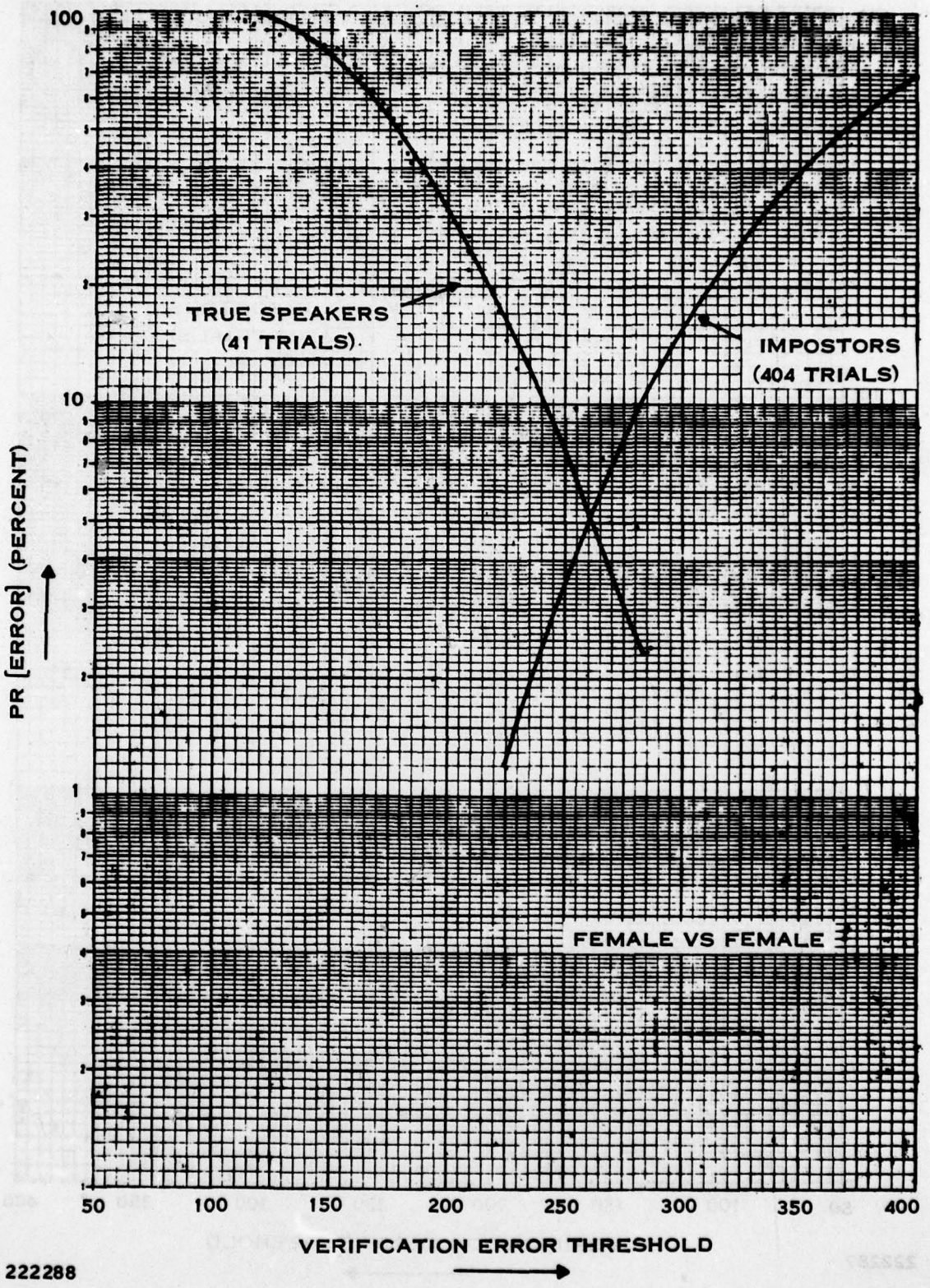


Figure 55. Female Speaker Verification Results

APPENDIX

APPLICATION OF CLUSTERING TO THE RECOGNITION OF SPOKEN CONNECTED DIGITS

SECTION VII

CONCLUSIONS AND RECOMMENDATIONS

The three major accomplishments on this program were (1) the installation of the Total Voice computer program on the ADM/BISS speaker-verification system at RADC, (2) the development of clustering techniques for use as a quantitative method of developing sets of reference patterns for speaker-independent word recognition, and (3) the collection of two large data bases for use in digit sequence recognition and speaker verification experiments.

The installation of the Total Voice computer program included several tasks that were precursors to the actual program installation:

1. Debugging and installation of a vector comparator for Euclidian distance computation
2. Debugging and installation of digital filter modifications
3. Development of a set of reference patterns for speaker-independent digit recognition.

The stated contract goal was to have less than 2 percent total error (rejection + substitution) on the recognition of six-digit sequences. This was achieved on a test set of 1,060 sequences from 106 subjects, but was not on another test set of 3,395 sequences from 11 subjects (see Section V). Improvement of speaker-independent connected digit recognition performance is one of the goals of other work now in progress at Texas Instruments. An obvious improvement to be suggested is the incorporation of results of that work into the Total Voice program.

Other suggested improvements to the speaker-independent digit recognition are to:

1. Increase the sample rate of the digital filter input in order to eliminate the aliasing of the sibilant energy
2. Postmultiply the filter outputs for proper amplitude normalization to ensure adequate sibilant and third formant energy
3. Use the maximum output of filters 14–16 instead of the average or sum to aid in detecting sibilants
4. Use T-function peaks to help anchor scanning error valleys at the actual reference points, especially for nasal-vowel transitions
5. Improve the reference scanning patterns to, for example, use multiple patterns for reference point No. 1 of "FOUR," and better account for nasal formants occurring in filter 2
6. Speed up the sequence recognition procedure by using dynamic programming plus simple substitution to satisfy check digit constraints.

Specific improvements to the clustering procedure would be to:

1. Account for the occurrence of "outliers" in the design data
2. Speed up the algorithm
3. Use much larger sets of design data.

Although there were insufficient data in the speaker-verification trials for very meaningful results, some recommendations can still be made. These include:

1. Normalization of speaker-specific recognition errors during verification by expected recognition errors for that speaker
2. Use of a sequential strategy for verification
3. Requiring lower errors for updating than for verification.

This last suggestion concerning the updating thresholds is coupled with the whole question of "learning with a teacher who makes mistakes" (for example, references 19–22), which is the case when a reference file is updated when a Type II error (impostor accepted) is made. This question requires further study.

The final recommendation is for a complete evaluation of speaker verification performance, both using the data base collected during this contract and in an actual operational environment.

REFERENCES

1. A.E. Rosenburg, "Automatic Speaker Verification: a Review," *Proceedings of the IEEE*, Vol. 64, No. 4, April 1976, pp. 475–487.

2. B.S. Atal, "Automatic Recognition of Speakers from Their Voices," *Proceedings of the IEEE*, Vol. 64, No. 4, April 1976, pp. 406–475.

3. G.R. Doddington, "Speaker Verification," RADC-TR-74-179, April 1974, 785135/5GI

4. G.R. Doddington and B.M. Hydrick, "Speaker Verification II," RADC-TR-75-274, September 1975, A018901.

5. "Summary: Phase I–Analysis and Testing; Automatic Speaker Verification System," Contract No. F19628-74-C-0030, 30 July 1974.

6. G.R. Doddington, "Personal Identity Verification Using Voice," *Electro 76 Professional Program*, paper 22-4, May 11–14, 1976.

7. G.R. Doddington, R.E. Helms, and B.M. Hydrick, "Speaker Verification III," RADC-TR-76-262, August 1976, B014720L.

8. B.G. Secrest and R.E. Helms, "Remote Terminal Speaker Verification," RADC-TR-77-169, May 1977, A040827.

9. G.E. Peterson and H.L. Barney, "Control Methods Used in a Study of the Vowels," *JASA*, Vol. 24, No. 2, pp. 175–184, March 1952 (Reprinted in *Readings in Acoustic Phonetics*, I. Lehiste, ed. Cambridge, Mass: The MIT Press, 1967).

10. D.R. Reddy, "Speech Recognition by Machine: A Review,"*Proceedings of the IEEE*, Vol. 64, No. 4, April 1976, pp. 501–531.

11. J.S. Kenyon and T.A. Knott, *A Pronouncing Dictionary of American English*, Springfield, Mass: G.&C. Merrian Co., 1953.

12. A. Fejfar, "Test Results–Advanced Development Models of BISS Identity Verification Equipment; Volume 1, Executive Summary," Mitre Technical Report, MTR-3442, Vol. 1, Rev. 1, September 1977.

13. M.R. Anderberg, *Cluster Analysis for Applications*, New York: Academic Press, 1973.

14. B. Everitt, *Cluster Analysis*, London: Heinemann Educational Books, Ltd., 1974.

15. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, New York: Wiley, 1973.

16. J.L. Flanagan, *Speech Analysis Synthesis and Perception*, Second Edition, (Berlin: Spinger-Verlag, 1972).

17. G. Fant, "Analysis and Synthesis of Speech Processes," Chapter 8 in *Manual of Phonetics*, ed. B. Malmberg (Amsterdam: North-Holland Publishing Co., 1968).

18. K. Fukunaga, *Introduction to Statistical Pattern Recognition* (New York: Academic Press, 1972).

APPLICATION OF CLUSTERING TO THE RECOGNITION OF SPOKEN CONNECTED DIGITS

Presented at

The Fourth International Joint Conference on Pattern Recognition

November 7-10, 1978
Kyoto, Japan

by

Robert L. Davis

TEXAS INSTRUMENTS
INCORPORATED

CENTRAL
RESEARCH
LABORATORIES

13500 NORTH CENTRAL EXPRESSWAY P O BOX 5936 DALLAS, TEXAS 75222

APPLICATION OF CLUSTERING TO THE RECOGNITION OF SPOKEN CONNECTED DIGITS

Robert L. Davis
Central Research Laboratories, Texas Instruments, Inc.
P.O. Box 225012, MS 05, Dallas, TX 75265, U.S.A.

This paper describes the use of hierarchical clustering in the selection of a set of reference patterns for use in speaker-independent connected digit recognition. The speech representation is the output of a 16-channel filter bank. A two step recognition process is used to solve the time alignment problem. First locate possible reference points at phoneme boundaries, and then use sequences of reference points to form a time aligned, down-sampled representation of the hypothesized digit. The clustering method is an agglomerative one (N to 1) followed by an iterative optimization on each of the final ten cluster sets. Samples of the clustering results are given both for patterns used in reference point location and for patterns used to select hypothesized digits. Results of recognition of 6-digit sequences are given for a test data set of 1060 sequences from a total of 106 speakers.

INTRODUCTION

The impetus for the recognition of spoken connected connected digits at Texas Instruments came not from a stand-alone application but rather to provide the claimed identification front-end to a voice authentication (speaker verification[1-3]) system. The procedure for authenticating a person's identity in an operational environment is: entry of the person into a lockable booth, inputting the claimed identity to the system via badge reader or keyboard, computer prompting of a verification phrase, speaking the verification phrase, and verification using the spoken phrase. Two important considerations in such an operational environment are maximum throughput and user acceptance (directly affecting the Type I or false dismissal error rate). As an aid to increasing both throughput and user acceptance, the idea of using speech to input as well as verify the claimed identity was introduced and called "Total Voice"[4]. In Total Voice, the badge reader, keyboard and prompting phrase are eliminated, and both identification and verification are done using a single spoken six-digit utterance, uniquely assigned to each subject.

SPEECH PROCESSING

The continuous speech data is sampled at time intervals of ten milliseconds duration and the amplitude is determined for each of 16 frequency bands listed in Table 1.

Table 1: Filter Bank Definition

Filter	Center Freq.	Bandwidth (6 dB)
1	280	250
2	395	280
3	525	310
4	630	340
5	750	360
6	900	360
7	1080	360
8	1265	365
9	1480	365
10	1725	365
11	1985	365
12	2285	360
13	2640	365
14	3150	625
15	3720	635
16	4235	615

This work was partially supported by Rome Air Development Center, Contract No. R30602-76-C-0329.

The values of filters 14-16 are averaged into one value, and data for the resulting 14 inputs is regressed using sine/cosine basis functions, normalized, and quantized to three bits. The value of the first sine and cosine regression coefficients and a normalized energy (quantized to 4 bits) are included with the 14 inputs to yield a 17 element vector for each time sample.

At this point it should be made clear that any direct comparison between a reference pattern formed by concatenating selected 17 element vectors and a similar pattern extracted directly from the input samples is futile due to the temporal inconsistencies of speech even from the same speaker. In order to solve this problem, registration points can be located in each word at points exhibiting a maximum spectral change and a representative "recognition" pattern for each word can be extracted from the input sequence by interpolating a fixed number of vectors between registration points. This input recognition pattern can then be compared to reference recognition patterns for either speaker independent or dependent word recognition or for speaker verification.

The next question is how to locate these registration points in the input waveform. One method would be to measure the spectral transitionitivity of the input by computing a difference between adjacent input vectors. However, the occurrence and size of such transitionitivity peaks are not reliable enough for locating registration points, although this function is a useful design tool. Instead, for each time sample, a "scanning" pattern is extracted from the input filter bank waveform. This scanning pattern is a five time-sample window of the input (five successive 17-element vectors concatenated together), centered on each time sample in turn, which is compared to a set of reference scanning patterns. Local minima (valleys) in the squared distance between input and reference patterns (scanning error) indicate potential registration points. The locations of these potential registration points are then paired with the locations of adjacent registration points in a dynamic programming routine. Minimum/maximum time differences between adjacent points are used as thresholds and the paired error between points is weighted by the deviation from the expected time difference.

The "best" sequences of potential registration points are then found and used to define recognition patterns from the input as shown in Figure 1 for the digit "six". The squared distances between the input and reference recognition patterns are then computed to provide a list of "hypothesized" words. At this point, constraints such as non-overlapping words, minimum error, or other syntactic constraints are imposed upon the sequence of words. As an example, in the case of the voice authentication front-end, the syntactic contraints are

1) six digits are in the sequence,
2) two of the digits are check digits (linear combinations of the other four),
3) certain digit pairs are disallowed, and
4) all digits must be different,

The total number of allowable sequences is thus limited to 320. More detail of the concepts in this section may be found in Doddington, et.al.[4].

An important point should be made here that this processing concept applies equally well to either isolated or continuous speech. Whereas, many other techniques used in word recognition are inherently restricted to isolated speech due to the need to detect end-point locations, no such restriction applies to the above techniques.

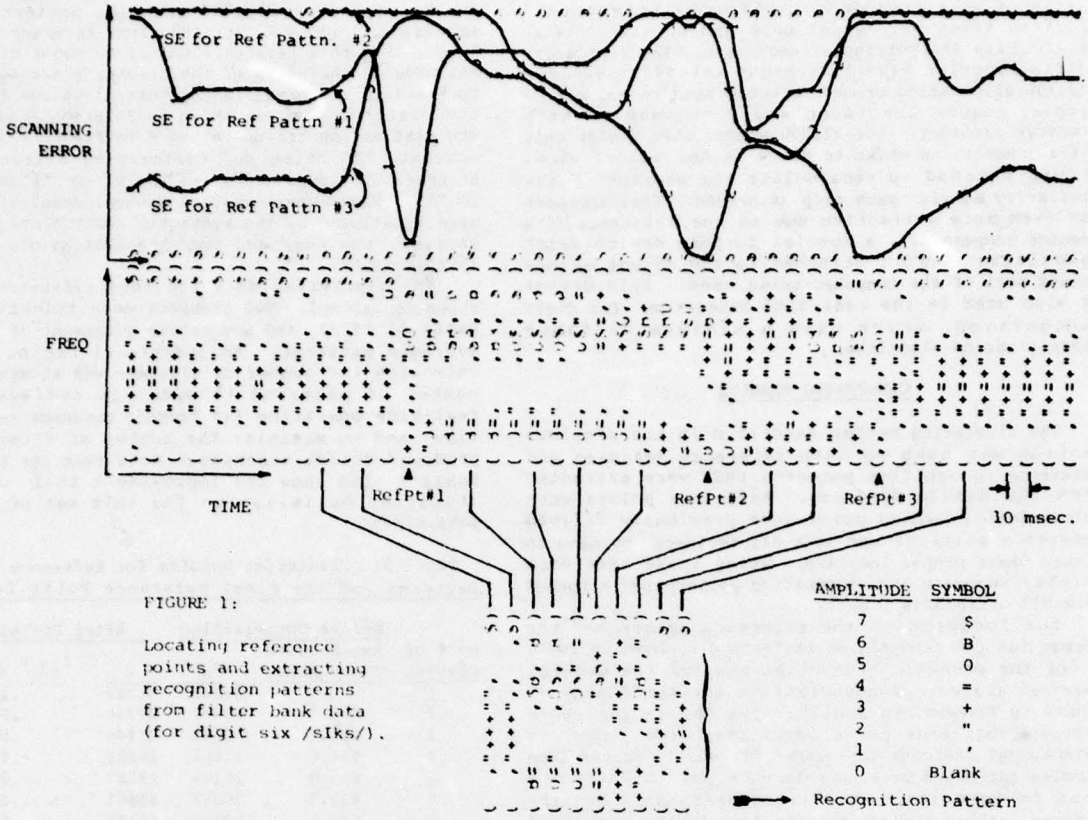


FIGURE 1:

Locating reference points and extracting recognition patterns from filter bank data (for digit six /sIks/).

CLUSTERING METHOD

Although some scheme such as using scanning and recognition patterns for time registering the input speech waveform is mandatory for speaker independent word recogntion, the drawbacks of using only one "representative" pattern for each reference point and for each word becomes readily apparent. The variations due to context, dialect, idiolect, and actual physical characteristics of the speaker must be accomodated by allowing multiple scanning and recognition patterns in order to gain acceptable performance levels.

Since the number of representative patterns needed was not known a priori, a hierarchical clustering procedure [5-7] was used on both pattern types extracted from the design data to determine sets of candidate clusters for each type of pattern. The design set consisted of two samples of each digit said in connected contexts from each of 85 subjects (42 males, 43 females.

The agglomerative method used in this study combined the two clusters having the smallest average distance between the points in the two clusters, i.e. combined the i and j clusters which have the minimum

$$\frac{1}{n_i n_j} \sum_{\vec{x} \in \chi_i} \sum_{\vec{x}' \in \chi_j} d(\vec{x}, \vec{x}')$$

where n_i = the number of $\vec{x}$'s in class χ_i
n_j = the number of $\vec{x}$'s in class χ_j and in this case $d(\vec{x}, \vec{x}') = \| \vec{x} - \vec{x}' \|^2$

The second step used was to iteratively improve upon the partitions from the hierarchial clustering by moving samples from one group to another if such a move improves the value of some criterion function. This step used the iterative optimization method in Duda and Hart[7], minimizing the sum-of-squared-error criterion, J_e, written as

$$J_e = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{\vec{x} \in \chi_i} \| \vec{x} - \vec{m}_i \|^2$$

At each step $\vec{x}$ should be transferred from class χ_i to χ_j and the means updated if

$$\frac{n_j}{n_j + 1} \| \hat{x} - \vec{m}_j \|^2 < \frac{n_i}{n_i - 1} \| \hat{x} - \vec{m}_i \|^2 \qquad (1)$$

Specifically $\hat{x}$ is moved to the class χ_j having the smallest $\frac{n_j}{n_j + 1} \| \hat{x} - \vec{m}_j \|^2$

An additional property of this selection for J_e is that a set of equally divided clusters is favored over a set containing both small and large clusters. This can be seen by considering $n_j << n_i$ in equation (1), which yields approximately,

$$\frac{n_j}{n_j + 1} \| \hat{x} - \vec{m}_j \|^2 < \| \hat{x} - \vec{m}_i \|^2$$

Thus, for $n_j = 1$, the distance $\| \hat{x} - \vec{m}_j \|^2$ need only be less than twice the distance $\| \hat{x} - \vec{m}_i \|^2$ to the old mean to be transferred to class χ_j.

The question still remains of choosing the "proper" number of clusters. One set of the iteratively optimized clusters was chosen on the basis of

1. minimum value of $(J_e^n - J_e^{n+1}) / J_e^n$
2. value of J_e (data sets with large J_e's favor using more clusters), and
3. subjective judgement as to when no new unique features are present in the average patterns for each of the clusters.

In actually implementing the hierarchical clustering portion of this program the question arose as to whether to calculate the similarity matrix (a

matrix of distances or correlations between all clusters taken pair wise) once and update it or to recalculate the entries at each step. The similarity matrix contains N(N+1)/2 unique entries. Since a sequential updating procedure would tend to accumulate errors, double precision would be used to help preserve accuracy. For 32,000 words, this yields only N=178. Hence, in order to allow larger values of N, it was decided to recalculate the entries of the similarity matrix each step as needed. This approach was even more attractive due to the existence of a "vector comparator", a special purpose device which calculates $\|\vec{x}-\vec{y}\|^2$, attached to the direct memory access port of the computer being used. This device is also used in the real-time processing for digit recognition, which uses a minimum distance classification algorithm.

CLUSTERING RESULTS

The clustering method described in the previous section was used on both reference scanning and reference recognition patterns that were extracted from the design data set. Reference points were automatically marked using some previously defined reference patterns and were all reviewed by hand to insure their proper location. Using these reference points, scanning and recognition patterns were formed from all acceptable digits.

The location of the reference points and the format for the recognition patterns are shown in Table 2 for the phonetic transcriptions for the General American dialect pronounciations for the digits as found in Kenyon and Knott[8]. The Δ's in the table indicate reference point locations; the number in parentheses indicate the number of equally spaced time samples extracted between reference points which are used to form the recognition patterns; and the numbers without the parentheses specify the number of time samples prior to the first reference point or after the last reference point at which the input spectral waveform is sampled for use in forming the recognition pattern. This is more clearly demonstrated by the recognition pattern for "six" shown extracted from an input spectral waveform in Figure 1.

Table 2: Reference Point Locations and Recognition Pattern Format Definitions for the Digits "Zero" through "Nine".

$z\ \Delta_{(4)}\ I\ r\ \Delta_{(4)}\ o\ \Lambda$	$f\ \Delta_{(6)}\ a\ I\ \Lambda\ v$
$w\ \Lambda_{(6)}\ \Lambda\ \Delta\ n$	$s\ \Lambda\ I\ \Lambda_{(4)}\ k\ \Lambda_{(2)}\ s$
$\Lambda\ \Lambda_{(2)}\ \Lambda\ u\ \Lambda_{(6)}\ \Lambda$	$s\ \Lambda\ \varepsilon\ \Lambda_{(4)}\ v\ \Lambda_{(4)}\ \Lambda\ n$
$\theta\ \Delta\ r\ i\ \Delta_{(6)}$	$\Lambda\ e\ \Delta_{(6)}\ t$
$f\ \Delta_{(6)}\ o\ \Lambda\ r$	$n\ \Lambda_{(6)}\ a\ \Delta\ I\ n$

The results of the clustering of reference scanning and recognition patterns produced 100 scanning patterns for the twenty-four reference points defined in Table 2 and 55 recognition patterns for the ten digits. A complete delineation of these patterns is given in Davis, et al[9]; however, a few patterns will next be reviewed to indicate the results of the clustering procedure. Generally, differences due to idiolects had no influence; context and speaker's sex determined clusters for the scanning patterns; dialects and speaker's sex determined clusters for the recognition patterns.

The first example of scanning pattern clusters appears in Figure 2, for the final reference point in "two". For this reference point, no major distinction was made on the basis of the speaker's sex due largely to the lack of spread in the formant values for /u/ on the basis of sex, as shown in Peterson and Barney[10]. The distinction between A and B in Figure 2 is due to context, "A" being /u/ followed by silence and "B" being /u/ followed by /s/ as in "6" or "7" or /z/ as in "0". Since vowel-vowel and vowel-nasal transitions were disallowed by the syntactic constraints described earlier, the /ue/ and /un/ transitions did not occur in this study.

The clustering data for this reference point is shown in Table 3. Two clusters were selected on the basis of $\Delta J_c/J_e$ and subjective judgement of the final averaged patterns. An additional factor used in selecting the number of clusters was to minimize the number of patterns in order to achieve almost real-time operation (<4 seconds sequence recognition time) and to minimize the number of false valleys produced during scanning. Note that the figures in Table 3 also show the improvement in J_c using the iterative optimization for this set of scanning patterns.

Table 3: Clustering Results for Reference Scanning Patterns for the Final Reference Point for "Two"

n= # of Classes	Before Optimization: Avg.Between Class Error	Before Optimization: J_e	After Optimization: J_c	After Optimization: $(J_e^n - J_c^{n+1})/J_e^n$
1	-	31242	31242	.272
2	476.3	22997	22736	.057
3	451.4	22169	21445	.058
4	446.6	21461	20205	.050
5	433.0	21174	19187	.064
6	421.3	20887	17961	.033
7	414.1	20314	17365	.027
8	389.5	20086	16902	.030
9	368.2	18914	16387	.023
10	367.0	18711	16008	-

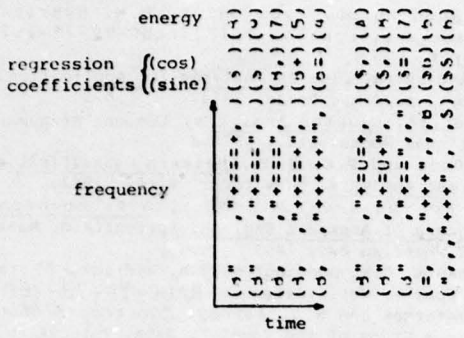
A second example of scanning pattern clusters appears in Figure 3, for the /sI/ transition in "six". This reference point is internal to the word and does not show the contextual variations present in the previous example. However, there was a significant difference on the basis of the speaker's sex as indicated by the formant locations in Figure 3.

Differences between the two male patterns, A and B, are due to energy profile, intensity of the third formant, and a slight variance in the reference point location. Differences between patterns C and D are due largely to energy profile. The distinguishing characteristic of pattern E is the absence of sibilant energy in the top filter since the upper cutoff of the top filter was below the sibilant frequency location for several females.

Figure 4 shows an example for the digit "nine" of the clustering results for recognition patterns. The differences among these patterns are due primarily to dialect and speaker's sex (Table 4) and in the formant locations in Figure 4. 75% of the speakers in the design set had recognition patterns for both samples of the digit nine in the same cluster.

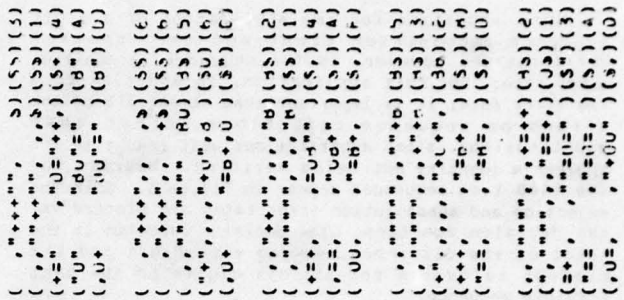
Table 4: Male/Female Distribution and Vowel Labels for Figure 4 Recognition Patterns

Pattern	Contributors to Patterns: Males	Contributors to Patterns: Females	Vowel/Diphthong
A	18	0	Λ^I or a^I
B	26	5	Λ^I
C	27	0	ae
D	13	10	ae
E	0	39	Λ^I
F	0	32	ae



No. of Contributors:		A.	B.
	Males:	41	43
	Females:	52	33

FIGURE 2: Reference scanning patterns for the /u-/or/us/ transition at the end of "two".



	A.	B.	C.	D.	
Males:	34	37	3	4	4
Females:	4	3	26	30	22

FIGURE 3: Reference scanning patterns for the /sI/ transition in "six".

Very few differences exist among the nasals. However, during the vowels, (1) different vowels occur, (2) a varying amount of diphthongization occurs, (3) a third formant is present in patterns A and C, (4) nasalization sometimes occurs as shown by the energy in the lowest filter in patterns B and E, and (5) a greater degree of positive spectral tilt occurs in patterns C, E and F as shown by the larger values of the cosine regression coefficient.

Although the location of primary and secondary schooling for most subjects in the design set was in Texas, a noticeable difference in this background occurred for pattern A vs. pattern C. For pattern A, about 40% of the subjects were schooled in Texas; about 40% in the middle or far West; and about 20% in New England. However, for pattern C about 70% of the subjects were schooled in Texas or the Southern U.S.; about 15% in the middle West but had lived many years in Texas; and only about 15% were relative newcomers from the middle West. This background is consistent with the presence of the diphthong in pattern A and only the vowel /ae/ in pattern C.

EVALUATION

In order to test the clustering results a test set of 106 speakers was used. Each speaker repeated one of ten possible sets of 6-digit sequences resulting in 1060 sequences. The test subjects were different from the training subjects. In these tests there is an inverse relationship between processing time and error rate. The evaluations used here were for "almost real-time" processing, comparing results using multiple reference patterns determined from the clustering program to those using only one representative pattern of each type.

Table 5 shows digit recognition results for both evaluation runs. These results are independent of all of the syntactic constraints mentioned earlier.

Table 5: Digit Recognition Results

Digit	# of Occurrences	Percent Correct Single Reference	Percent Correct Multiple Reference
0	689	89.3	93.6
1	434	93.8	93.5
2	626	71.4	79.6
3	738	83.9	89.3
4	711	81.9	89.0
5	796	96.6	99.0
6	764	91.4	96.7
7	756	69.0	86.2
8	405	82.2	98.0
9	441	68.3	89.0

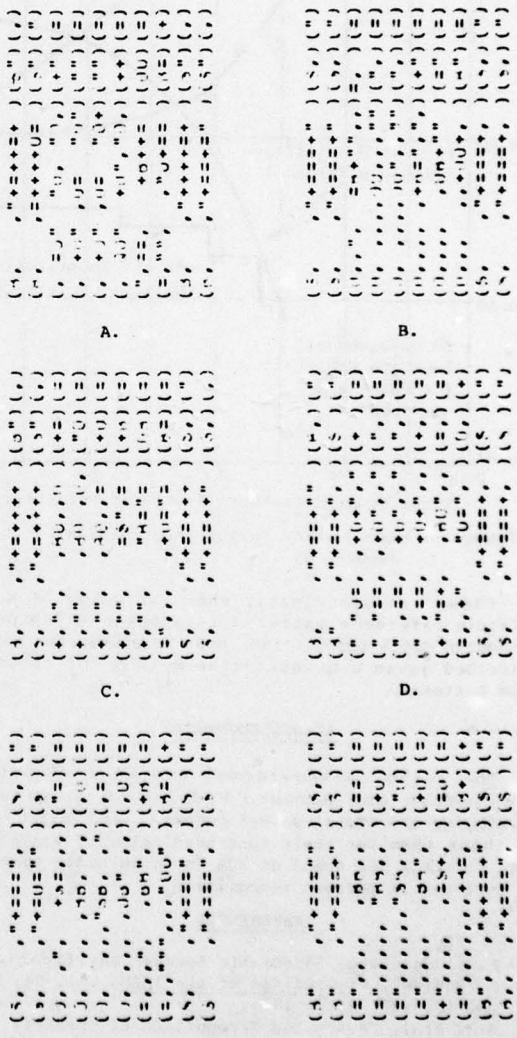


Figure 4: Reference recognition patterns for "Nine"

More important for the application of a digit sequence recognition front-end for speaker verification, however, is the sequence recognition error rate. For this application, in addition to a low error rate, it is important that almost all of the errors be sequence rejections rather than substitutions, since substitutions will result in the speaker's identity not being verified. Results for the 1060 test sequences appear in Figure 5. Both the rejection and substitution error rates are plotted vs. the decision function. The decision function is the total of the distances between each digit and its closest reference for all six digits of the best possible sequence.

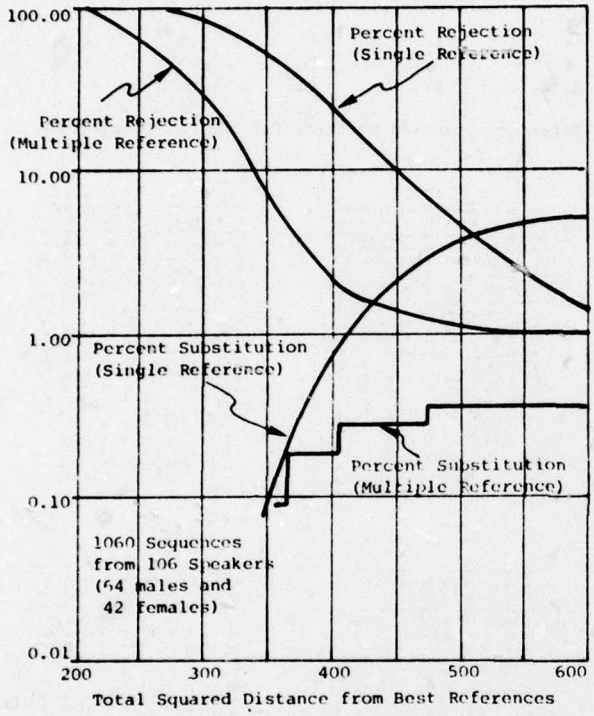


Figure 5: Recognition Performance for Six-Digit Sequences

These results clearly show the value of using multiple reference patterns in speaker-independent connected digit recognition, and the clustering method described gives a quantitative method for choosing such patterns.

ACKNOWLEDGMENTS

The author acknowledges the work of George Doddington and Barbara Hydrick in originally developing the "Total Voice" concept, and would like to thank them for their continued help. I would also like to thank Ken Abend of RCA for originally sparking my interest in pattern recognition.

REFERENCES

1. A.E. Rosenburg, "Automatic Speaker Verification": a "Review", Proceedings of the IEEE, Vol. 64, No.4 pp. 475-487, April 1976.
2. B.S. Atal, "Automatic Recognition of Speakers from Their Voices", Proceedings of the IEEE, Vol. 64, No. 4, pp. 406-475, April 1976.
3. G.R. Doddington, "Personal Identity Verification Using Voice", Electro 76 Professional Program, paper 22-4, May 11-14, 1976
4. G.R. Doddington, R.E. Helms, B.M. Hydrick, "Speaker Verification III", RADC-TR-76-262, August 1976.
5. M.R. Anderberg, Cluster Analysis for Applications, New York: Academic Press, 1973.
6. B. Everitt, Cluster Analysis, London: Heinemann Educational Books, Ltd., 1974.
7. R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, New York: Wiley, 1973.
8. J.S. Kenyon and T.A. Knott, A Pronouncing Dictionary of American English, Springfield, Mass: G. & C. Merrian Co., 1953.
9. R.L. Davis, G.R. Doddington, B.M. Hydrick, "Total Voice Speaker Verification", RADC-TR-78-260.
10. G.E. Peterson and H.L. Barney, "Control Methods Used in a Study of the Vowels", JASA, Vol. 24, No. 2, pp. 175-184, March 1952 (Reprinted in Readings in Acoustic Phonetics, I. Lehiste, ed. Cambridge, Mass: The MIT Press, 1967).